DOCUMENT RESUME

ED 261 505 EC 180 574

AUTHOR Stone, C. Addison; And Others
TITLE Assessment and Remediation of Complex Reasoning in Specific Subgroups of Learning Disabled Adolescents. Final Report.
INSTITUTION Northwestern Univ., Evanston, Ill. Dept. of Communicative Disorders.
SPONS AGENCY Department of Education, Washington, DC.
PUB DATE Jul 84
GRANT G008102719
NOTE 208p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS *Abstract Reasoning; Age Differences; Case Studies; *Cognitive Processes; *Intervention; *Learning Disabilities; Problem Solving; Secondary Education; *Student Characteristics; Teaching Methods

ABSTRACT

The study is described which examined quantitative and qualitative differences among learning disabled (LD) subgroups and between LD and normal Ss in reasoning and problem solving behaviors. The research strategy involved (1) detailed analyses of the behavior of subgroups of LD adolescents and of matched normal achieving adolescents in a task requiring the use of complex reasoning skills; (2) a detailed follow-up of the progress made by individuals exhibiting specific reasoning difficulties over a series of individually designed instructional sessions; and (3) the development of materials to help LD practitioners diagnose and remediate reasoning deficits in adolescents. Three LD subgroups were identified and their performances compared with three non-LD control groups: no discrepancy (ND) LD, low verbal (VL) LD, and low performance (LP) LD. Comparisons were made among subgroups and with same age and younger controls. The intervention phase was designed to explore the utility of a single-subject design and a Piagetian clinical interview strategy for remediating reasoning and problem solving demands. Among findings were that the three LD subgroups demonstrated differential performance on the bending rods task, LD Ss performed more like ninth grade than fourth grade controls, and performance of ND and LV Ss was mixed. It was concluded that not all LD adolescents have difficulties with reasoning and problem solving skills, and that the specific nature of the dificulties varies as a function of the type of LD and of the task demands. (CL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED261505

Assessment and Remediation of Complex Reasoning in Specific Subgroups of Learning Disabled Adolescents

C. Addison Stone
Ellice A. Forman
Caroline J. Anderson
Francine Matthews
Jennifer Rupert
Beye Fyfe

Learning Disabilities Program
Department of Communicative Disorders
Northwestern University

Final Report

Department of Education Grant No. G008102719

July 1984

Any opinions, findings, and conclusions expressed in this report are those of the authors and do not necessarily reflect the views of the Department of Education. Address correspondence to either of the first two authors: Learning Disabilities Center, Northwestern University, 2299 Sheridan Road, Evanston, IL 60201.

EC180574

BEST COPY AVAILABLE

Acknowledgments

This project could not have been accomplished without the assistance of a number of people. The participation of the administration, staff and students from the following Chicago Metropolitan area school districts is gratefully acknowledged: Glenbrook, Deerfield, Maine Township, Leyden Township, Elmwood Park and Oak Park. We would also like to thank the students and staff of the Cove School in Evanston for their help. We are indebted to a number of current and former students from the Program in Learning Disabilities at Northwestern University for their assistance in data collection, coding and analysis. In particular, we would like to thank Joan Brubaker, Lydia Conca, Pat Croasemun, Barbara Curl, Leslie DiMario, Myra Kraker, Marcia LaPorte, Donna Michals, Debbe Scheffel, Susan Seifert, Patricia Shimon, Barbara Snyder, Susan Weingartner, and Beverly Whitmire. We also depended upon the typing, editing, and organizational skills of Mary Rooney, Ruth Schultz and Lou Detlefsen. Finally, we would like to thank the following people who served as consultants on the project: Jane Blalock, Dave Cordray, Eleanor Duckworth, Jack Easley, Doris Johnson, Mieko Kamii, Deanna Kuhn and Tom Weaver.

Table of Contents

Page

Introduction
- Statement of the Problem and Rationale 1
- Objectives of the Project 3
- Overview of the Report 4

Literature Review
- Overview 5
- Performance of Normal Adolescents in IV Task Settings 5
- Reasoning and Problem Solving in LD Adolescents 7
- Piagetian Studies of Selected Exceptional Populations 11
- The Delay vs Difference Issue 19
- Utility of the V-P Discrepancy in Predicting Task Performance 20

Assessment Phase
- Purpose and Overview 25
- Documentation of Need for Research on Thinking Skills
  - Teacher Interviews 25
  - Data from Teacher Questionnaires 28
- General Procedures
  - Subjects 28
  - Task Materials and Administration Procedures 32
  - Data Coding 34
- Group Differences in Use of the IV Strategy
  - Overview of Analyses and Technical Considerations 40
  - Group Differences in Task Success
    - Comparison Among Three LD Subgroups 41
    - Comparisons With Same Age and Younger Controls 42
    - Comparisons With Equivalent and Lower IQ Groups 44
    - Summary 46
  - Comparison of Performance on the Rods and Conductivity Tasks . . 47
  - Comparison of Rods and Conductivity Tasks to Woodcock-Johnson . . 49
- Group Differences in General Task Approach
  - Overview of Analyses and Technical Considerations 50
  - Group Differences in Task Approach Among Three LD Subgroups . . . 51
  - Comparisons With Same Age and Younger Controls 52
  - Summary 56
- Development of Teacher/Clinician Rating Scale 57

Intervention Phase
- Purpose and Overview 60
- Methods 62
- Case Studies of Progress Across Sessions 67
- Summary and Preliminary Suggestions for Interventions 90

General Discussion and Conclusions 93
- Differential Performance of the Three LD Subgroups 94
- Comparisons to Normal Controls 95
- The Nature of Reasoning and Problem Solving Difficulties 96
- Educational and Theoretical Implications 99
- Limitations of the Present Study & Need for Future Research 100

References 103

Tables and Figures 114

Appendices
- A. Teacher Questionnaire 184
- B. Synopsis of Coding System 186
- C. Teacher/Clinician Rating Scale Manual 192
- D. Dissemination 201

INTRODUCTION

Statement of the Problem and Rationale

The primary purpose of the current project was the documentation of quantitative and qualitative differences among learning-disabled (LD) subgroups and between LD and normal subjects in reasoning and problem solving behaviors. A secondary goal was to explore ways in which our procedures and findings could be used by LD practitioners. The ultimate goal of research such as this would be the development of principles and procedures for the identification and remediation of reasoning and problem solving difficulties in LD adolescents. The general research strategy involved (1) detailed analyses of the behavior of subgroups of LD adolescents and of matched normal-achieving adolescents in a task requiring the use of complex reasoning skills; (2) a detailed follow-up of the progress made by individuals exhibiting specific reasoning difficulties over a series of individually-designed instructional sessions, and (3) the development of materials to help LD practitioners diagnose and remediate reasoning deficits in adolescents.

Complex reasoning plays a central role in adolescent functioning in both academic and non-academic settings. The conception of abstract reasoning and systematic problem solving skills, and, in particular, Piaget's concept of formal operations (Inhelder & Piaget, 1958), has had an important impact on educational theory, policy, and practice during the past twenty years (Ausubel & Ausubel, 1966; Larson & Dittmann, 1975; Kuhn, 1979; Hurd, 1978; Lovell & Shayer, 1978). Psychologists and educators have argued that the development of new reasoning and problem solving skills in early adolescence is necessary for the mastery of the abstract ideas and critical thinking central to higher education (Kuhn, 1979; Hurd, 1978; Peel, 1971). More recently, the importance of these concepts for our understanding of the cognitive demands of non-educational settings has also been explored (Kuhn & Ho, 1977; Kuhn, 1979; Linn, 1978a; Kuhn & Brannock, 1977; Capon & Kuhn, 1979; Erwin & Kuhn, 1979).

Given the central role of reasoning and problem solving skills in our understanding of adolescent development, principles and procedures for dealing with deficiencies in these skills are greatly needed to provide comprehensive services to LD adolescents. Unfortunately, we know very little about reasoning and problem solving skills in this population since traditional assessment frameworks and testing batteries do not include sufficient coverage of such skills. Due to the lack of knowledge about the reasoning and problem solving potential of LD adolescents, contradictory assumptions about their abilities are often made. In some instances, learning-disabled students are steered away from science, math, and other academic-track courses because of their demands for complex reasoning and systematic problem approach. On other occasions, it is assumed that reading and writing are the only barriers that LD students face in these courses. In fact, there is very little empirical evidence on which to base either assumption. A detailed study of reasoning and problem solving skills among specific groups of LD adolescents should help provide the evidence needed for accurate curriculum planning and occupational advising. Furthermore, a study designed to monitor closely the long-term changes in reasoning and problem approach shown by LD adolescents during remediation will serve as a model for future attempts to enhance these skills in this group. Finally, by including analyses designed to assess directly

(1) the benefits for assessment and remediation of the information obtained and (2) the ease with which LD specialists can be prepared to make use of it, the potential impact of the project was assured.

While recent reviews of research on adolescent reasoning and problem solving have led to the conclusion that there is, in fact, a qualitative change in the manner in which children approach problems as they move into the early adolescent years (Neimark, 1975; Day, 1978), there is increasing evidence (Martorano, 1977; Shayer, 1979) that the change is not one of a global cognitive reorganization, as was originally proposed (Inhelder & Piaget, 1958). This fact has led several authors to discourage discussion of a general stage transition and to call for more detailed studies of particular reasoning and problem solving skills (Day, 1978; Keating, 1980; Neimark, 1979; 1980). Consistent with these trends, the current project focused on a single skill, the isolation of variables in a multivariate context.

The isolation-of-variables strategy was chosen for two reasons. First, the strategy has face validity. It represents a critical thinking skill of clear utility in a wide range of day-to-day situations (e.g., finding the cause of an allergic reaction) to which all adolescents, LD and non-LD, are exposed (Kuhn, 1979). Thus, its status in LD adolescents is of some concern independent of more general issues. Second, the strategy has a rich research tradition. In their pioneering work on adolescent reasoning, Inhelder and Piaget (1958) argued that the systematic isolating of variables was a key criterion in the assessment of the transition to the stage of formal operations. For this reason, the isolation-of-variables strategy has been the focus of a large number of assessment and training studies with normal children and adolescents (Day, 1978; Martorano & Zentall, 1980; Wollman, 1977; Stone & Day, 1978; Linn & Levine, 1978; Bredderman, 1973; Lawson & Wollman, 1976; Kuhn & Angelev, 1976; Kuhn, Ho, & Adams, 1979). Thus, it was possible to capitalize on a large body of knowledge about the determinants of strategy use both in designing the assessment and remediation phases of the study and in interpreting the data.

Several additional features of the current research served to maximize the utility of the information obtained. (1) The use of subgroups of LD adolescents with differing profiles of disabilities provided more detailed knowledge of subject characteristics than would have been possible with a heterogeneous group, while still providing a more representative sampling of the LD population than is possible in the study of a single LD subgroup. (2) Detailed videotape analyses of task behaviors helped to focus attention on the reasoning and problem solving process and facilitated the description of specific deficits. (3) The assessment of the target strategy in two separate contexts and the inclusion of a third measure of reasoning skills added to the generalizability of the findings. (4) The comparison of the LD groups with normal-achieving control groups matched for age and IQ served to highlight those aspects of the reasoning process unique to the LD groups. (5) Finally, the use of detailed behavioral observations across a series of individually-tailored intervention sessions highlighted the potential benefits of intervention for specific reasoning and problem solving difficulties.

Objectives of the Project

The general goals of the project can be expressed in a series of specific objectives. Each of these objectives is discussed briefly here as it was originally conceived.

In the final section of the Report (Discussion and Conclusions), the relevance of the project findings to these objective is discussed.

1. To determine the relative frequency of spontaneous use of the isolation-of-variables strategy:
   a. among specific subgroups of LD adolescents;
   b. between each LD subgroup and normal-achieving adolescents.

2. To determine the relative ease with which the reasoning strategy can be elicited from those LD and normal-achieving adolescents who fail to use it spontaneously.

   Since past research has shown that the majority of those normal-achieving adolescents who do not use the isolation-of-variables strategy spontaneously can readily be induced to do so in a second administration of the task by presenting a series of intervening structured probe questions, it was important to determine how many of the LD adolescents in the sample exhibited a similar "elicitable" use of the strategy.

3. To determine the generality of the problems seen in the initial tasks.

   In order to gain some assurance that the problems seen in specific individuals are not restricted solely to the particular task used for the assessment, two additional reasoning tasks were administered in a second session. The first of these additional tasks was another measure of the isolation-of-variables strategy. The second was a standardized task which requires some of the same subskills as an isolation-of-variables task but which requires the subject to use them in a different context.

4. To determine the specific subcomponents of the isolation-of-variables strategy which cause difficulties for the LD adolescents.

   In order to obtain a more detailed understanding of the reasoning problems of LD adolescents, it was necessary to analyze their behaviors and verbalizations during each step they took to solve the task.

5. To determine if, and how, the difficulties encountered by the LD adolescents in the assessment tasks differ from those encountered by younger, normal-achieving pre-adolescents.

   This information aided in our understanding of the extent to which the reasoning problems encountered by the LD adolescents represented a deviation from the typical developmental progression.

6. To determine the overlap between the reasoning problems identified in the assessment phase and those most evident to classroom teachers and LD specialists in clinic and school settings.

While formal assessment of reasoning skills is rare, LD specialists may have implicit frameworks for recognizing difficulties in reasoning. It was important to explore the utility of such frameworks in the context to be studied and to determine how the conclusions reached by practitioners differ from those obtained with detailed coding techniques. This information was important in assessing the utility of the research findings and in bridging the gap between research and practice.

7. To develop a modified assessment procedure for identifying reasoning problems which can be used by LD clinicians in the field.

   The ability of LD clinicians to identify specific reasoning problems with the use of a behavior rating scale based on the detailed coding procedures provided information concerning the most useful means of translating the information obtained from the first phase of the research into clinical practice.

8. To determine if a series of remediation sessions can be effective in improving deficient reasoning skills in LD adolescents.

   The study of adolescents' progress across the series of remediation sessions can further refine our understanding of the severity of the reasoning problems isolated during the assessment phase. The degree of progress and the amount of skill transfer provide an estimate of the utility of remediating reasoning deficiencies.

9. To identify successful intervention units for the reasoning problems seen.

   This information will lead to the identification of realistic goals for remediation.

10. To determine the relative amount of progress which can be made in mastering the isolation-of-variables strategy among the LD adolescents exhibiting different reasoning difficulties during the assessment tasks.

    This information can help to refine the assessment information by further highlighting those aspects of reasoning difficulties which are unique to specific groups of LD adolescents.

11. To develop guidelines for use by LD practitioners in translating assessment information into remediation goals and techniques.

    The products of this objective will include a document describing principles and procedures for assessment and remediation of reasoning problems and a demonstration film including narrated examples taken from the videotapes collected. These materials should be of utility in the professional training of practitioners.

Overview of the Report

This report is divided into four major sections. The first section contains a review of existing literature related to the theoretical and empirical issues central to the present project. The second and third sections contain a detailed description of the procedures and findings from the project. The

second section is devoted to the assessment phase of the project and contains a report of findings relevant to the first seven objectives. Included in this section are reports of the relative incidence of success at using the isolation of variables strategy across LD and normal-achieving subgroups and descriptive analyses of differential task approach. The development of a teacher rating scale for assessing reasoning and problem solving is also described. The third section is devoted to the intervention phase of the project and covers objectives 8-11. Included here are case studies of interventions and a discussion of possible intervention strategies. The final section of the report contains general conclusions.

LITERATURE REVIEW

Overview

The following literature review is more extensive than is typical for a research report. Since the approach taken in this project to the study of LD adolescents is a relatively new one, we feel that the breadth and level of detail is needed to provide a context for the research. The first section of the review covers existing research on the performance of normal-achieving preadolescents and adolescents on tasks similar to those used in this project. The second section reviews prior research on reasoning and problem solving in LD adolescents. Included here is a discussion of the pilot research which led to the development of this project. The third section of the review contains a discussion of previous research with LD and MR populations conducted from a Piagetian perspective. This section serves to highlight certain theoretical and empirical issues which have arisen in past research. One of the most important of these issues is that of developmental delay vs. difference. This issue is addressed in more detail in the fourth section of the review. Finally, since the LD subgroups studied in the present project were defined in part on the basis of discrepancies in verbal and nonverbal intelligence, the utility of this discrepancy for clinical and research purposes is discussed in a final section of the review.

Performance of Normal Adolescents in IV Task Settings

Assessment Studies

The present section will be limited to those studies of normal adolescents directly relevant to establishing the context for the proposed research. Thus, it will focus directly on the strategy of isolating variables in a multivariate context.

Inhelder and Piaget (1958) were the first researchers to assert that the isolation-of-variables strategy was a developmental acquisition unavailable to pre-adolescents. Since the account of their research first appeared, researchers have shown a steadily growing interest in children's and adolescents' use of the strategy. Until the late 1970s, studies using the strategy were largely attempts to test Piaget's theory of formal operations. These studies used Inhelder and Piaget's (1958) original procedures and scoring criteria to examine the incidence of the strategy in different age groups and/or the relationship of performance in isolation-of-variables tasks to performance in other formal operations tasks (Lovell, 1961; Jackson, 1965;

Dulit, 1972; Somerville, 1974). As it became clear that, contrary to Piaget's theory, the spontaneous use of the strategy was far from universal (Dulit, 1972), researchers attempted to teach children and adolescents to use the strategy as a means of assessing Piaget's assertion that the strategy could not be taught until a child had developed the necessary cognitive prerequisites (Siegler, Liebert, & Liebert, 1973: Lawson, Blake, & Nordland, 1975; Lawson & Wollman, 1976; Case, 1972; Case & Fry, 1973).

These studies produced two major findings. First, while the specific ages varied (presumably as a function of differences in procedures and scoring criteria), those studies with more than one age-group demonstrated an interaction between age and benefit from instruction (Lawson & Wollman, 1976; Case, 1972). Second, those researchers who included transfer tasks found that the trained skills showed very little evidence of generalizing to tasks requiring other formal operations strategies (Ross et al., 1976; Lawson et al., 1975; Lawson & Wollman, 1976).

The result of this line of investigation was a growing disenchantment with Piaget's notion of a general stage of formal operations (Neimark, 1975; Blasi & Hoeffel, 1974; Keating, 1980; Linn, 1978a, Stone, 1977). As a result of this disenchantment, researchers have focused their interest on the specific reasoning strategies identified by Inhelder and Piaget (1958) in order to develop a better characterization of adolescent reasoning skills. In this context, the isolation-of-variables strategy has become the focus of research for its own sake (Danner & Day, 1977; Stone, 1977; Stone & Day, 1978; Stone & Day, 1980; Linn & Levine, 1978; Pulos & Linn, in press; Wollman, 1977; Martorano & Zentall, 1980).

These recent studies have yielded two important findings about how adolescents use a specific strategy. First, the data indicate that for some adolescents, use of the strategy is closely related to task or procedural variations (Linn, 1978b; Stone & Day, 1978). Furthermore, the examiner can elicit the strategy with only minimal prompting from subjects who fail to use it spontaneously (Danner & Day, 1977; Stone, 1977; Stone & Day, 1978; Kuhn, Ho & Adams, 1979), and additional tests indicate that the elicited strategy-use is genuine (Stone & Day, 1978; Kuhn Ho & Adams, 1979; Neimark, 1980). Second, in contrast to earlier findings with respect to the general stage notion, ready access to the isolation-of-variables strategy appears to be universal among normal adolescents by age 14 (Stone & Day, 1978; Neimark, 1979; Stone, 1980).

Studies of the isolation-of-variables strategy with the normal population thus provide solid evidence of new developments in reasoning skills in adolescence which appear to be universal. In light of these findings, the status of such skills in LD adolescents becomes a significant issue for research.

Multiple Session Intervention Studies

Piagetian-inspired research on the use of the isolation of variables strategy has focused on intervention as well as on assessment issues. The bulk of the intervention literature is not relevant to the research project summarized here and will not be discussed. However, one new direction in this research literature merits some attention. Kuhn and Phelps (1979) have argued that training studies that employ short-term interventions and experimental designs are incapable of providing crucial information about the nature of the

developmental process. This is true for several reasons: the existence of change is assessed after-the-fact through the use of pre-post scores; the developmental process that was induced may bear little resemblance to that which occurs in more naturalistic circumstances; the training procedures may induce superficial modeling behaviors in the subjects rather than genuine, irreversible cognitive growth. Instead, they propose using a multiple-session, observational design with subjects who exhibit, on the pretest, no evidence of the ability to isolate variables. Over a period of time, subjects are asked to solve problems that require the use of the IV strategy. The examiners do not explicitly teach the IV strategy nor do they reinforce the subjects' behaviors. However, feedback about the effectiveness of their problem-solving strategies is provided by the materials. Evidence of developmental change is inferred from the careful observation of subjects' reasoning and data gathering strategies over time.

This methodological strategy, although new, has been successfully employed in a few studies using IV tasks and normal subjects (Kuhn and Phelps, 1982; Forman, 1981). The information it can yield about the nature of the learning process in LD subjects has potential value for both assessment and remediation. For example, if LD adolescents learn to master the IV strategy at a rate and in a manner similar to that of younger normal-achieving children, then instructional activities appropriate for younger children should be employed. If, however, LD adolescents show evidence of differences in initial and subsequent task approach, then unique educational programs may be needed to stimulate their thinking.

Reasoning and Problem-Solving in LD Adolescents

While there has been a long and continuing research interest in reasoning and problem-solving skills among younger LD children (Strauss & Werner, 1942; Strauss & Kephart, 1955; Klees & Lebrun, 1972; Inhelder, 1976; de Ajuriaguerra et al., 1976; Blalock, 1977; Meltzer, 1978), very little attention has been paid to these issues in the adolescent population. This neglect is particularly surprising in light of recent research with normal adolescents and the reports of clinicians and educators who work with LD adolescents.

The reports of clinicians and educators have long suggested that the reasoning skills of the LD adolescent population are grossly inadequate to meet the demands of the high school setting or of the "real world" (Deshler,1978). Wilcox (1970) noted that LD adolescents may evidence breaks in the continuity of thought, poor organization, difficulty in selecting alternatives, and an inability to sustain attention. Siegel (1974) noted that the LD adolescent is often "disorganized" and has an "inability to plan systematically and to follow through." Kronick (1978) has argued for a greater appreciation of the potential for difficulties in social reasoning.

Two research studies confirm in part the observations of clinicians and serve to highlight several important issues in need of further research. Havertape (1976; Havertape & Kass, 1978) asked a group of LD adolescents and a group of same-aged normal adolescents to talk aloud as they read and attempted to solve a series of thirteen tasks. The tasks consisted of problems of four types: (1) simple arithmetic word problems (price comparisons), (2) the completion of number series, (3) the solution of "word puzzles", and (4) the writing of a limerick. Havertape coded verbalizations into several categories of relative

problem-solving sophistication under the general headings of Getting the Information (reading the problem correctly), Understanding the Problem, and Solving the Problem (using logical and efficient steps). There were significant differences between the normal and LD groups within all three categories. Havertape and Kass concluded that, while some of the differences stemmed from basic deficiencies in reading, writing, and math skills, there was evidence that "in many cases learning disabled students have no attack strategies to apply to problem solution; or, if they do, they do not effectively use them." (Havertape & Kass, 1978, p.98).

A recent unpublished study by Skrtic (1980) provides some additional evidence of an unsystematic approach to problem-solving among LD adolescents. In the context of a larger study of the math difficulties of LD adolescents, Skrtic (1980) administered a group measure of reasoning skills to LD and normal-achieving adolescents matched for age, sex, and school (IQ scores for the controls were not available). The measure, developed by Lawson (1978) consisted of a series of twelve questions based on several of Inhelder and Piaget's (1958) formal operations tasks. The adolescents observed demonstrations of Inhelder's tasks, responded to written questions about each demonstration, and wrote justifications of their answers. The answers and justifications were used to generate a binary score for each question, yielding a total possible score of twelve. Although there was considerable overlap between the two distributions, the LD group ($x = 2.4$) scored significantly lower than the control group (4.8). While Skrtic cautioned that the significant difference might be in part attributable to differences in IQ, he interpreted his general findings as evidence of a delay in cognitive development in his LD sample.

While these two studies appear to provide some support for the clinical reports that LD adolescents often evidence difficulties in complex reasoning, several aspects of their findings lead to questions in need of research. First, while both studies included somewhat heterogeneous samples of LD adolescents by virtue of their sampling procedures, and while both include evidence of significant intra-group variation, neither study provides information concerning the relative frequency of difficulties in adolescents evidencing different specific disabilities. A second question, closely related to the first, concerns the specific nature of the difficulties encountered by the LD adolescents. While Havertape & Kass are able to tell us that some LD adolescents use fewer "logical and efficient steps" to solution, they note that detailed information about the kinds of strategies used and their sequence of use would be of enormous benefit in designing effective remediation programs. Preliminary data on such a remediation program are presented by Arbitman-Smith and Haywood (see below) Also, as Havertape and Kass imply, we need more detailed information about the problem-solving process to determine which of the LD adolescent's difficulties are secondary to academic deficiencies and which are primary deficits. Clearly, more research is needed in this area before definitive conclusions can be drawn which address these questions.

A recent study by Arbitman-Smith and Haywood (1980) utilized a program called Instrumental Enrichment; a teaching model developed by Feuerstein et al. (1980), designed to enhance the growth of deficient cognitive skills. Instrumental Enrichment (IE) consists of a systematic framework for mediated learning experiences utilizing fifteen teaching instruments, each focused upon

a specific, deficient cognitive function, and designed to facilitate appropriate generalization of principles and strategies (i.e., evaluation of relevant information, planning strategies, comparison and interpretation of results, etc.). These strategies are normally assumed to develop spontaneously through learning experiences in the environment mediated by adults. The program constitutes approximately 300 hours of instruction over a period of at least two years. In several studies utilizing IE across the U.S., the samples were not strictly LD, but consisted of various exceptional groups including LD, EMR, BD, those in Resource Rooms, and Mexican-American (second language) slow learners (x IQ = 80.42). In preliminary data from one study with fifth and sixth grade identified LD students in Nashville, no significant difference was noted between IE and comparison groups during post-testing, and no significant gains were recorded within the IE group for pre- and post-testing. However, mastery progress testing (administered on the same day as post-testing, consisting of utilizing portions of standardized tests which were similar in principle or strategy to the IE tasks) revealed various levels of transfer for the IE group. The transfer noted was reflected in increased attention to detail, improved approach-to-task strategy, and increased persistence, as well as improved intrinsic motivation, and positive behavior changes noted by significant others (teachers and/or parents). From this it was concluded that there do exist basic cognitive skills which can be taught to LD students through a non-categorical program such as Instrumental Enrichment, skills which can then be subsequently transferred to new tasks. The question raised was how such cognitive changes can best be measured to appropriately reflect newly developed abilities.

Pilot Research

The pilot data which led to the current project provided some information about the reasoning skills of LD adolescents (Stone, 1981). The data also served to motivate several methodological features of the proposed research.

A series of three isolation-of-variables tasks was administered to a heterogeneous group of LD adolescents as part of a comprehensive diagnostic evaluation. The tasks and procedures were similar to those used in a previous study with normal children and adolescents (Stone & Day, 1980). All three tasks involved a set of ten rods varying in length, material, and diameter, a stand into which the rods could be placed, two at a time, and a pair of identical weights which could be attached to the rods in order to assess their relative bending.

Task 1 assessed the subject's spontaneous use of the isolation-of-variables strategy. The examiner used one pair of rods to demonstrate to the subject that "some rods bend more than others" and asked the subject to use pairs of rods to "find out for sure what makes a difference for bending." While the subject worked, the examiner took notes on whether the two rods used in each test constituted an unconfounded test (i.e., varied only on the variable being tested). A subject's score for this task was the number of variables out of four (length, material, diameter, and place of weight attachment) which were consistently tested in an unconfounded manner across the task as a while.

Task 3 was identical to Task 1 (except that the set of rods consisted of different instances of the variables) and was administered in order to assess improvements in performance as a function of the experience gained during

Tasks 1 and 2. Task 2 consisted of a series of questions which served as a means of encouraging the subject to focus on the difference between a confounded and an unconfounded test (i.e., the presence of a second, confounding variable). Similar cuing has been successful in improving the performance of normal adolescents who fail to use the strategy spontaneously (Stone & Day, 1978; 1980).

The pilot subjects consisted of a heterogeneous sample of LD adolescents ($\underline{N}$ = 36) ranging in age from 12 to 19 years ($\bar{X}$ = 14.8). The diagnosis of learning disabilities was predicated on normal verbal or nonverbal intelligence (85 or above on one Wechsler subscale), freedom from primary sensory deficits or emotional disturbance, adequate educational opportunity, a significant discrepancy between ability and achievement in one or more areas (including oral language, reading, writing, math, and visual-spatial skills), and evidence of specific deficiencies in basic information processing. The mean fullscale IQ for the sample was 100.1 (range = 80-123). A wide range of specific disabilities was represented.

The major results of the study are easily summarized. First, approximately one-half of the sample (20 out of 36) used the strategy spontaneously (on Task 1) and an additional 20% (8/36) could use the strategy after minimal prompting (on Task 3). Thus, this reasoning skill was available to the majority of LD adolescents in the sample. It is important to note, however, that strategy-status was not independent of primary area of disability. Of the 20 adolescents with primary difficulties in reading or in written language, 19 showed evidence of the strategy by Task 3, and only 1 failed to use the strategy on any task. In contrast, of the 7 adolescents with primary disabilities in math and/or visual-spatial skills, 4 of the 7 failed to show any evidence of the strategy. Similarly, 3 of the 5 subjects with primary disabilities involving language comprehension also failed to use the strategy.

Of most importance is the fact that these instances of reasoning difficulties do not appear to be attributable to normal developmental or individual differences. The strategy-absent students were not younger than their peers, as one would expect from past research with normal subjects (Stone & Day, 1978). Also the failure to use the strategy was not directly related to IQ. Finally, certain qualitative features of the behavior of the strategy-absent subjects were not evident in their normal, strategy-absent peers. The LD adolescents were more likely (1) to identify fewer of the potential variables without prompting from the examiner, (2) to attend to inappropriate details or unlikely variables (e.g., the age of the wooden rods), (3) to make multiple confounded tests of the same variable in succession, and (4) to begin Task 1 using single rods rather than pairs. Also, two behaviors common in normal strategy-absent subjects were not observed in the LD strategy-absent group: (1) using a single pair to draw conclusions about two variables and (2) testing two identical rods to ascertain that they bend the same.

These unique behaviors, as well as the other features of the pilot data, provide some indication that an isolation-of-variables task setting is a useful context in which to study reasoning difficulties in LD adolescent populations, but a more careful and detailed study was needed to answer several questions raised by the findings. First, a closer look at the reasoning skills of specific subgroups of LD adolescents seemed warranted. More information was needed about the relationship between reasoning problems

and specific learning disabilities. Second, comparisons with carefully matched control groups were necessary to highlight the severity and unique nature of the reasoning problems of the LD population. Third, a more detailed analysis of each adolescent's behaviors in the task was necessary to isolate the nature of the problems encountered. Also, it was important to determine whether a more careful analysis of other features of the reasoning process (such as the nature of the conclusions drawn or the efficiency of the testing sequence) would in fact reveal difficulties among those LD adolescents who appear to have a command of the isolation-of-variables strategy when measured with a global summary score. Fourth, information was needed about the benefits LD strategy-absent adolescents could gain from more structured and long-term intervention in order to assess their potential for new learning. These issues were addressed in the present research project.

Piagetian Studies of Selected Exceptional Populations

Mentally Retarded Children

In reviewing the literature on Piagetian tasks and mental retardation there are several problems that must be kept in mind. First,"retardation" is defined differently by various investigators. Although one researcher might consider a retarded person as anyone with an IQ below 75, Inhelder (1968), for example, used retardation to refer only to the group whose IQ lies (approximately) between 50 and 75, but not to those who IQ is below 50. IQ scores are sometimes but not always used to specify range of deficiency, and even IQ scores vary according to the measure used. Finally, any categorization is arbitrary, (someone with IQ 75 is "retarded" and someone with IQ 76 is "not retarded").

A second preliminary consideration is the fact that people are retarded for a wide variety of reasons including brain damage, psychological disability, and genetic disorder. These differences in etiology result in differences in behavior which may or may not relate to a person's performance on Piagetian tasks. Few investigators have controlled for these factors.

Finally, it should be pointed out that while "Piagetian tasks" are compared from one study to another but there may be substantial differences in the way the actual experiments are conducted and scored.

With these reservations in mind, this review will proceed with a discussion of Inhelder's investigations because her studies formed the basis from which much of the research with exceptional children has arisen.

The pioneering studies of mental retardation from a Piagetian perspective were conducted by Inhelder and reported in her book The Diagnosis of Reasoning in the Mentally Retarded (1968). Her study was based on clinical interviews of 159 subjects who had been diagnosed as mentally retarded by teachers, psychologists, or physicians. The subjects were of mixed etiology, they ranged in age from 7 to 52 (all but 4 under 25), and they had IQs ranging primarily between 50 and 90. The experimental tasks were largely taken from Piaget's conservation of matter, weight, and volume. Tasks used were: 1) clay, 2) Dissolution of sugar, 3) Boxes - same weight, different sizes. Inhelder's major conclusions can be summarized as follows:

1. 90% of the mentally retarded subjects reasoned in a way consistent with Piagetian theory. The developmental sequence was the same as in normals, although the speed of development was slowed and the transition from one stage to another seemed to progress more slowly than is noted with normals.

2. Deficient populations fixate at lower levels than do normal populations. Inhelder suggested the following comparisons (Inhelder, 1968, p. 292-3).

	Binet M.A.Level	Piagetian Stage
Idiot	0-2 year	Sensory-motor, prior to language
Imbecile	2-7 year	Instinctive, no operations
Retardate	7-12 year	Concrete operations
Slow learner		Formal operations, eventually

In particular, Inhelder said, "To be retarded means, therefore: to be able to think by concrete operations, but not by formal operations." (Inhelder, 1968, p. 294). She noted that children 12 to 13 years old who were at the borderline level of retardation were characterized by "a fixation at the level of concrete thought . . . We do not find even the beginnings of formal operations in any of these subjects. In fact, as soon as we present these backward children with problems whose solutions require a formal level of organization - for example, a combinatory system - they do not behave like normal preadolescents of 10 or 11, but like young children of 6 or 7 who could be performing the most elementary concrete operations. Thus, it seems that as soon as the problem becomes too complex, the mentally retarded child gives up the idea of trying to organize the situation and simply repeats the same actions over and over again in the hopes an accumulation of repetitions will sooner or later, in some way have the desired effect." (Inhelder, 1966, p.312-313). Inhelder used the term "false equilibrium" to label the ceiling level of the retarded (see later discussion).

3. Some cases (10%) showed abnormal oscillations from one stage to the next. Inhelder identified these oscillations of being of three types.

   a. Progressive reasoning - occurs in slow learners only and is characterized by progress in learning during the test or interview situation. For example, a subject may be incapable of analyzing a new problem but can reason step by step once an introduction is provided. Inhelder hypothesizes that the reasoning of these subjects is initially blocked by social and affection factors.

   b. True oscillations - defined as constant fluctuations between two levels and are caused by such affective factors as anxiety, suggestibility, and hesitation.

   c. Retrogressive reasoning - refers to progressive deterioration of reasoning. That is, a child may affirm certain beliefs initially, but then later question and abandon them. Inhelder believes that cases of retrogressive reasoning arise when the subject has learned but not internalized a kind of reasoning. She believes that the subject's final stage is closer to his real intellectual level.

To summarize, Inhelder basically found that retardates follow the same sequence of stages as normals, they progress more slowly, and they reach a ceiling at a lower level than normals. In addition, some qualitative differences in reasoning were noted. Investigators since Inhelder have largely sought to confirm her finds.

Following Inhelder's lead, many investigators have attempted to confirm the finding that retardates follow the Piagetian sequence of cognitive development. Several reviews of the literature have been published showing general confirmation (e.g., Wilton and Boersma, 1974; Sternlight, 1981). The most extensive reviews are provided by Weisz and Zigler (1979) and Weisz and Yeates (1981).

Weisz and Zigler (1979) reviewed thirty-one studies to test the "similar sequence hypothesis, which by their definition, "holds that during development retarded and nonretarded persons traverse the same stages in precisely the same order and differ only in rate of development and ultimate ceiling they obtain." These authors considered studies at all levels: sensori-motor, pre-operational, concrete operational, and formal operational and divided the studies into two groups: 1) cross-sectional and order of difficulty or 2) longitudinal. They found that only five of the 31 studies did not provide consistent support for the similar sequence hypothesis. They felt that the exceptions were minor and that support for the hypothesis was convincing. However, they also saw the need for better research that involved, for example, directly comparing retarded and nonretarded subjects and that would control for organic etiologies.

The review by Weisz and Yates (1981) follows from concerns raised by the previous review. Specifically, the authors ask the following questions: When a mentally retarded child and a younger nonretarded child happen to be at the same level of intellectual development and at the same mental age, how similar are the two children in the processes by which they reason? The authors review material to test the "similar structure hypothesis" which suggests that if the structure of reasoning is the same in retardates and non-retardates, then people with the same M.A. level should pass Piagetian tasks at the same level.

The authors reviewed 30 different published experiments involving 104 comparisons of retarded and non-retarded subjects matched for M.A. Tasks involved preoperational and concrete operational tasks in many different areas such as moral judgment, spatial perception, seriation, and conservation. Of the 104 comparisons, the following results were obtained:

4% of the studies showed retardates performing higher than M.A. matched nonretardates;

24% non-retardates performed higher than M.A. matched retardates.

72% showed no significant difference between M.A. matched retardates and non-retardates, i.e., consistent with similar structure hypothesis.

Of 33 studies which excluded organics, only 10% were inconsistent with a similar structure hypothesis. Using a statistical procedure which could include all 104 comparisons, the distribution of the etiology-uncontrolled

studies was significantly different than expected by chance, favoring the non-retarded. Among those studies in which organics were excluded, there was not a significant difference from chance. This is consistent with the similar structure hypothesis.

In summary, recent review articles suggest that the majority of research done since Inhelder's original studies find results which confirm her statement that retarded persons progress along the same cognitive stages as normal persons and reason in the same way at equivalent M.A. levels.

As mentioned earlier, Inhelder came to define retardation as being "able to think by concrete operations but not by formal operations (Inhelder, 1968, p. 294). Many reviewers have recalled this definition and have sought to find studies which might demonstrate even a single case in which a retarded person showed formal operational thought (e.g., Neimark, 1980). There are, to our knowledge, no unequivocal examples. The series of studies most often cited to indicate lack of formal operations in mental retardation is the Stephens series mentioned below. A controversial series showing "formal operations" is also described below.

Stephens and her colleagues (Stephens and McLaughlin, 1974; Stephens, 1977) conducted a longitudinal study comparing 75 retarded (IQ = 50-75) and 75 non-retarded subjects ranging in age from 6 to 20 and involving three test periods, each testing separated by two years. In addition to WISC or WAIS scores, a series of Piagetian tasks was administered including these categories: conservation, logic, classification, operativity and symbolic imagery, and formal operations. The authors found that development continued in retarded subjects beyond the age of twenty, although it proceeded at a decelerating tempo. No subjects were successful in formal operational tasks (at least as far as data are presently available).

Lister (1970, 1972) did a series of studies which suggested that volume conservation could be taught to retarded subjects. Since volume conservation is considered by some to be a formal task (see Inhelder, 1968, p. 293), these studies are often mentioned by reviewers discussing whether or not retardates can reach formal operations. (e.g., Wilton and Boersma, 1974; Neimark, 1980). The Lister study used 30 mildly retarded subjects who were matched for IQ, CA, and pretest conservation. Experimental subjects were given interior and displacement volume training individually for 30 minutes. One and two week posttests were administered and an additional transfer test was given four weeks later. The control group was posttested. All experimental groups showed conservation of substance, weight, and volume. The control group showed no conservation at the initial pretest, but showed conservation after training.

One source of confusion about the results arises because Lister reported that 34 of her 104 retarded studies showed volume conservation even on the pretest. Volume conservation was noted in children with MAs as low as 5-7. This suggests that the task was not formal operational in the sense of Piaget and Inhelder. (Kahn, 1976; Neimark, 1980; Wilton and Boersma, 1974).

Kahn (1976) attempted to replicate Lister's study using 60 retarded (IQ = 55-75) and 60 non-retarded males, ages ranging from 12 to 16. Each group was divided into low and middle socio-economic level. Kahn pretested

with seven Piagetian measures of reasoning including conservation of substance and volume (clay balls) conservation of interior volume, oscillation of a pendulum, and equilibrium of a balance. The subjects were posttested two weeks later on the same tasks. The experimental group received training by the same procedure used by Lister. Kahn's results show that on the posttest at least eight of the thirty retarded subjects showed conservation on at least one volume task. All of the successful students were in the low SES group. Kahn concludes that his results confirm Lister's findings, but suggest then, that his subjects should be considered to be misdiagnosed as retarded (rather than concluding that the retarded can achieve formal operations).

In summary, in those few cases where formal operations have been observed in retarded subjects or taught to retarded subjects, there are alternative explanations which might be given. As yet there is no conclusive evidence of formal operations thought in retardates.

Aside from determining whether or not retarded subjects follow the Piagetian sequence and whether they can reach the formal operational level, what are the qualitative factors which describe a retarded person's approach to reasoning?

Inhelder (1966) described in some detail a difference that occurs between normals and retardates in achieving equilibrium. With the normal child, each element becomes consolidated with the others to form an equilibrium situation. In this way the system is then integrated into the next larger system. It is the "germ of further development." For example, concrete operations when completed can then be integrated into the system of formal operations. With the retarded child there does not seem to be such a firm consolidation. Instead, Inhelder refers to a "false equilibrium" in which there is an apparent stability but yet the system falls to a much lower level when challenged by more complex material (as when a retarded child is asked to do a formal task). The equilibrium is not stable enough to lead to a high level of reasoning.

Stephens (1977), in her longitudinal study of mental retardates mentioned earlier, also notes differences between the reasoning of retardates as opposed to the normals. She found that although the retardates follow the same sequence as normals, they do not achieve "the flexibility of thought, the ability to classify and reclassify, to group and subgroup information, to engage in the reversability of thought that is required in tasks of classification and conservation with the ease, dispatch, or thoroughness that was reflected in the performance of normals." This description of the retardate's reasoning is not inconsistent with Inhelder's concept of false equilibrium.

Schmid-Kitsikis (1976) studied the performance of the retarded and retarded psychotic on several Piagetian tasks in which she included conditions in which the task was slowed down (e.g., ball changed to sausage in several small steps), so she could observe the process of reasoning more easily. She found, as expected, that the retarded children showed the Piagetian hierarchy. She was also able to say that the retarded children proceeded with the task in a "normal" way--i.e., used successive discoveries to learn, had fixed goals, noticed errors, etc. Reid (1978, 1981), in reviewing the Piagetian studies of exceptional children reaffirms Schmid-Kitsikis' conclusions. She notes that retarded children are weak in deductive reasoning skills but show adequate

regulation of activity to achieve fairly stable equilibrium in a Piagetian sense.

In summary, a few studies have looked at the details of processing in retarded compared with non-retarded subjects, and both similarities and differences have been noted. It would be interesting to see further comparisons of the transitional stages in retarded vs. non-retarded subjects. Since retardates seem to disintegrate at the transition to formal operations, this would be an especially interesting period to examine in detail.

Learning Disabled Children

The Piagetian tradition offers a potentially rich theoretical and empirical framework for the study of problem solving behavior. However, the potential relevance of this framework for the study of childhood exceptionalities in general, and learning disabilities in particular, has been the subject of recent debate (Reid, 1978; Fincham, 1982; McFarland & Grant, 1982). One major issue concerns whether or not Piaget's general competence theory can be used to make predictions, and therefore, normative judgments, concerning the performance of individual subjects in specific task contexts. In the following discussion, we will sidestep this issue and adopt a pragmatic point of view. To the extent that comparisons of the performance of NA and LD children in Piagetian tasks yields interesting descriptions of differential task approach, the enterprise has value, regardless of its ultimate interpretation within a Piagetian framework.

The studies to be reviewed can be divided into two broad categories: those which report group differences in task success and those which provide more detailed measures or descriptions of differential performance. The latter will be given more attention.

By far the largest number of studies in the first category consists of comparisons of LD and NA subjects on one or more of Piaget's concrete operational tasks. These studies are evenly divided with respect to finding differences between LD and NA subjects. Five studies report significant differences (Andersson, Richards, & Hallahan, 1980; Johnston & Ramstad, 1983; Klees & Lebrun, 1972; Knight-Arest & Reid, 1979; Silvius, 1974), while five report no differences (Copeland & Weissbrod, 1983; Fincham, 1979; James, 1975; Kahmi, 1981; Meltzer, 1978). Differences in the number and nature of the specific tasks used preclude a detailed analysis of these studies. Furthermore, since these studies present only information about relative levels of task success, they offer no detailed information about qualitative aspects of problem solving behavior. Thus, we will not review all of them in detail.

Perhaps the most carefully executed study to report no difference in relative task success between LD and NA subjects is that of Meltzer (1978). The author compared the performance of a group of 35 LD children, aged 8 to 10 years to that of a group of 35 NA children. The LD subjects were attending a private school for the learning disabled and were reported as manifesting "visual perceptual problems." The children were administered eleven concrete operational tasks, consisting of multiple assessments of conservation of liquid quantity, conservation of number, seriation, and class inclusion. Both a correct judgment and an adequate explanation were required to pass a given

item. In approximately two-thirds of the matched pairs, both subjects either passed or failed the criteria, and there were no significant differences between the two groups on any of the tasks. It should be noted, however, that Meltzer dropped from her sample an unspecified number of LD subjects who failed to demonstrate an understanding of task-relevant concepts (e.g., "more", "some").

One of the most carefully executed studies to report significant differences between NA and LD children is that of Silvius (1974). Silvius compared the performance of three groups of 5-8 year-old children matched for fullscale IQ: two LD subgroups ($\underline{N}$ = 9 in each) and an NA group (N = 9) on the Concept Assessment Kit-Conservation, Forms A and C (Goldschmid & Bentler, 1968). Each LD subject had a discrepancy of 15 points or more between verbal and performance IQ, and the two subgroups were defined in terms of which subscale was higher. Silvius found a significant difference among the three groups on the total scores, with the two LD groups scoring lower than the NA group, but no differently from each other.

Additional analyses revealed that the two LD subgroups exhibited a different pattern of performance across the various conservation domains, with conservation of length being the most difficult for the low verbal group and conservation of area yielding the lowest incidence of conserving responses for the low performance group. These findings suggest that some LD children do indeed evidence difficulty in conservation activities, and that these difficulties may be related to the nature of the underlying disability. In this context, it is worth noting that the "visual perceptual" disabilities of the children studied by Meltzer may represent a less serious impediment to successful performance than the disabilities present in Silvius' subgroups.

Only two studies report data on relative levels of performance on formal operations tasks (Skrtic 1980; Stone, 1981). These studies are discussed in the section on Reasoning and Problem Solving in LD Adolescents.

As a group, the studies of relative levels of success on concrete and formal operational tasks still provide equivocal information concerning the problem solving skills of LD individuals. While many report significantly different success rates, some do not. A review of the handful of studies which provides qualitative analyses or descriptions of the behavior of LD subjects in Piagetian tasks offers more potentially useful information concerning problem solving skills in LD children.

The earliest reports containing qualitative descriptions of LD children's performance on Piagetian tasks come from European researchers. The subgroup which has received the most attention from these workers is the language disordered. Inhelder and Siotis (1963; reprinted in translation in Morehead & Morehead, 1976) report two case studies of language disordered children. The authors report that both of these children are typical of the majority of language disordered children they have tested in that they appear to have a global deficit in representational ability, particularly in nonverbal visual imagery. The authors also report that despite this deficit, these children were able to succeed at most concrete operational tasks. However, their approach to these tasks often involved compensatory strategies used to substitute action patterns for (ordinarily) imagined transformations. Very similar observations concerning language disordered children were made by

deAjuriaguerra and his colleagues at the same symposium (deAjuriaguerra, Jaeggi, Guinard, Kocher, Maquard, Paunier, Quinodoz, & Siotis, 1963). In a two-year follow-up of their original subjects, deAjuriaguerra and his colleagues also report that some of these children were experiencing serious difficulties in formal operational tasks as well (deAjuriaguerra, Jaeggi, Guinard, Kocher, Maquard, Roth, & Schmid, 1965; reprinted in translation in Morehead & Morehead, 1976). Additional confirmation of the problems with imagery skills in language disordered children can be found in more recent and systematic studies (Johnston & Ramstad, 1983; Kahmi, 1981; Sigel, McGillicuddy-DeLisi, Flaugher, & Rock, 1983). In addition, Klees and Lebrun (1972) report similar problems in a sample of dyslexics.

The issue of unique characteristics in the approach of LD individuals to Piagetian tasks which is raised by the work discussed above has been pursued by Reid and her colleagues, who have studied the performance of LD children using modified procedures developed by Genevan researchers to highlight the learning process rather than the level of attainment. Knight-Arest and Reid (1979; Reid, 1981) examined the influence of peer interaction as a "catalyst" for conservation acquisition in LD and NA children. The study involved three phases. In a pretest, all subjects were administered a standard conservation of liquid quantity task. Then, groups of three children (two conservers and one nonconserver) participated in a "party" in which the nonconserver was asked to share juice with the other two children, whose glasses were different shapes. In the third phase, conservation of liquid quantity was again assessed individually, at one hour and at two months after the "party." There were ten-party groups consisting of LD children, and ten with NA children. The LD children (N = 18) were 9-15 years old, from a private school for LD children, and had problems in reading, math, and/or visual-motor skills. The control group (N = 36) consisted of NA children aged 7 to 9 years. The authors found that 9 of the 10 initial nonconservers in each sample conserved on the immediate post-test. There was evidence, however, that the improvement shown by the LD children was superficial. Four of the 9 LD children who improved failed to conserve on the delayed post-test, compared to only 1 of the NA children. More importantly the authors report that the NA children often offered justifications for conservation which they had not heard during the party, but only one LD child did so. Reid (1981) reports some informal observations which lead her to the conclusion that the LD children "expected to find answers in the empirical aspects of the objects," and "resisted making inferences about their own actions" (p. 343).

A second study by Reid and Knight-Arest (1979; Reid, Knight-Arest, & Hresko, 1981) provides additional evidence of qualitative differences in the task approach of LD and NA children. Ten NA and 10 LD boys aged 10-12 were videotaped as they attempted to balance a series of wooden blocks, some of which had concealed counterweights. The LD boys evidenced many of the behaviors of younger, NA children (Karmiloff-Smith & Inhelder, 1975), such as making empirical adjustments to their initially random placements of the blocks. However, there was also informal evidence that the LD boys did not make efficient use of their own ideas to guide their actions. They often provided implausible, ad hoc explanations which were at variance with the evidence before them and showed signs of difficulty in organizing activity.

In summary, the studies within the Piagetian tradition provide a useful contribution to our knowledge about the problem solving skills of LD

individuals. There is evidence that many LD children experience difficulties with Piagetian tasks, and that these difficulties may vary as a function of the type of learning disability. (See Schmid-Kitsikis, 1969, for some case studies which further illustrate this claim.) Two major characteristics of the task approach of LD individuals emerge from the studies available to date. First, there is evidence that some children with language disorders make use of overt action patterns in order to compensate for deficient visual imagery skills. Second, some LD children fail to generate new information systematically in problem solving situations (block balancing or exploration of flexibility) and fail to make good use of the data they generate in developing or revising explanations. Both of these findings merit more careful study.

The Delay vs. Difference Issue

Deeply embedded in the history of the field of learning disabilities is the notion that learning disabled (LD) children exhibit a perceptual, linguistic, or cognitive organization which is qualitatively different from that of their normal-achieving peers. This assumption was a natural outgrowth of the dual roots of the field in the medical tradition, with its focus on neurological disorders and lesions, and in gestalt psychology, with its focus on organized systems. In recent years, this assumption has met with growing scepticism, largely because of an increasing focus on behavioral and educational skills. At present, the issue has crystallized into what is often called the "delay vs. difference" issue. At its base, this issue revolves around whether or not one believes the problems exhibited by an LD child are due simply to the fact that the child has the "learning readiness" of a younger child.

Research focused on the delay vs. difference issue is scarce, in part, because it is not at all clear how to conduct an adequate test. The complexities are nicely exemplified in the literature on children with language disorders. Reading recent reviews of the research in this area (e.g., Johnston, 1982; Kirchner & Skarakis-Doyle, 1983; Waryas & Crowe, 1982) suggests that whether or not one concludes that language disordered children are developmentally delayed or different depends in part on one's conception of language development. Researchers who view language as a bundle of discrete skills evaluate specific indices of linguistic achievement (e.g., plural morphemes, agent-object relations) in isolation from other aspects of the child's language functioning. This practice can lead to the conclusion that a child who performs poorly on a particular language measure is functioning in the same manner as a younger child who scores similarly on that specific measure. In contrast, a researcher who focuses on language as a tightly organized system looks for evidence of different relationships among linguistic subsystems in a child who shows depressed performance, and therefore is likely to view the child's language as qualitatively different from that of a younger child (Kirchner & Skarakis, 1983).

The delay vs. difference issue has received relatively little attention from researchers interested in the cognitive abilities of LD children. The few existing studies vary widely in theoretical orientation and methodology, and in the nature of the populations studied. For example, Sara Tarver and her colleagues, working from a discrete process approach, have conducted a series of studies focused on individual attentional and metacognitive processes in heterogeneous samples of school-identified LD children and adolescents (e.g.,

Tarver, Hallahan, Kauffman, & Ball, 1976; Tarver & Maggiore, 1979). These studies include careful measurement techniques and, of necessity, tightly structured task settings. The results indicate that the LD children perform significantly lower than their normal-achieving peers on specific measures, but that most of these differences are greatly attenuated, or actually eliminated, by early adolescence (Tarver & Maggiore, 1979). The authors use these findings to argue for a delay model.

In contrast, researchers within the Piagetian tradition can be interpreted as arguing (often only implicitly) for a difference model (e.g., Inhelder, 1976; Klees & Lebrun, 1972; Reid, Knight-Arrest & Hresko, 1981; Schmid-Kitsikis, 1973). These researchers tend to use complex tasks, demanding a broad range of skills, and to study clinic-derived samples of a relatively homogeneous composition. In general, findings suggest that while many LD children exhibit a level of task success similar to that of younger normal children, their approach is different. There is also evidence of compensatory strategies which represent a combination of skills not usually seen in normal children. Most of the studies, however, do not provide a detailed data base from which definitive conclusions can be drawn.

Utility of WISC-R Verbal-Performance Discrepancies in Predicting Task Performance

The Wechsler IQ scales are widely used for evaluating intelligence at every age level from preschool to adult. One of the features that makes these scales popular is that each provides a Verbal IQ score and a Performance IQ score in addition to an overall or Fullscale IQ score. Because of this division into verbal and nonverbal (performance) scores, clinicians are able to evaluate an individual's relative abilities. However, the division has led to controversy about the meaning of a discrepancy between verbal and performance scores. Since the publication of the first Wechsler scale more than forty years ago (Wechsler, 1939), researchers and clinicians have noted that some individuals and diagnostic groups seem to show unusually large discrepancies in one direction or another. The result is that numerous studies have been published with evidence and counter-evidence about the clinical significance of the V-P discrepancy. There is a group of studies which claim, for example, that delinquents have higher performance than verbal scores on the WISC or WAIS (e.g., Camp, 1966; Prentice and Kelly, 1964; Henning and Levy, 1967; Andrew, 1974). There is another group of studies that looked for V-P discrepancies in emotionally disturbed or schizophrenic populations (e.g., Dean, 1977; McHugh, 1963; Wechsler and Jaros, 1965; Schoonover and Hertel, 1970). Still another group of studies looked for the V-P discrepancy as a sign of brain damage (e.g., Beck and Lam, 1955; Hopkins, 1964; Holroyd and Wright, 1965; Black, 1974).

Although still popular, these kinds of studies of the V-P discrepancy are controversial because the results are inconsistent, there is considerable overlap of various diagnostic groups showing similar V-P discrepancies, and too little attention has been given to normal data (Matarazzo, 1972; Zimmerman and Woo Sam, 1972; Kaufman, 1979). Kaufman summarizes his review of the subject as follows:

. . . virtually the entire V-P literature--not just the studies pertaining to brain dysfunction--is beset by contradictions and a lack of success in

identifying characteristic patterns for various groups. Poorly defined samples or samples that fail to control for essential variables are probably partially responsible for the inconsistencies. However, another likely source of the problem is the fact that a V-P discrepancy may signify quite different things for different individuals. Also, some V-P discrepancies may be misleading and not indicative of different verbal and nonverbal skills or they may be totally meaningless. Finally, some of the contradictory research results may be due to the magnitude of the V-P discrepancies for normal individuals (Kaufman, 1979, p.25).

The field of learning disabilities has met the same kind of controversy and inconsistency as other clinical areas when investigating the V-P discrepancy. These studies will be reviewed in detail, but first data from the WISC standardization sample will be summarized to provide a context for discussing the V-P discrepancy in a special population.

V-P Discrepancy and the Normal Population

The discrepancy data from the 2200 individuals in the standardization samples were analyzed by Seashore (1951) for the WISC and Kaufman (1976) for the WISC-R. Kaufman's data, much like Seashore's, indicated that 43% of the sample showed a discrepancy of at least 10 points and 24% showed a discrepancy of at least 15 points in either direction. Twelve percent of the standardization sample had at least a 20 point discrepancy and less than 1% had a discrepancy of 34 points or higher.

The mean WISC-R discrepancy of the standardization sample was 9.7 points. The mean discrepancy was the same for all age groups and did not differ significantly by race or sex. A slightly lower mean discrepancy (9 points) was found in children of unskilled workers compared to the discrepancy (11 points) for children of professionals. The discrepancy for children whose Fullscale IQ was below 80 was slightly lower than the discrepancy for higher scoring children.

In the standardization sample there was an equal distribution of higher verbal and higher performance scores. However, children of professionals were more likely to have higher verbal than performance scores if there was a discrepancy, and this was not true in children of unskilled laborers.

Kaufman (1979) notes that a 15 point discrepancy on the WISC-R is frequently noted as abnormal by clinicians, but in fact it occurs in nearly one-fourth of the normal population.

LD and the V-P Discrepancy

Studies which have investigated WISC-R V-P discrepancies in the learning disabled or reading impaired can be divided into two basic categories: (1)those that are simply concerned with the size of the discrepancy and (2) those concerned with the direction of the discrepancy (e.g., Are low verbal-high performance scores more likely to be associated with learning disabilities than high verbal-low performance scores?). Within these two categories there are also two basic ways to investigate the relationship between V-P discrepancy and LD: (1) groups can be determined on the basis of learning characteristics and then compared for WISC scores; (2) WISC scores

can be used to assign groups which are then compared for learning characteristics. The size discrepancy issue will be reviewed first looking at learning characteristic groups and then WISC groups. Then the discrepancy direction issue will be reviewed again looking at learning characteristic groups and then WISC groups.

V-P Discrepancy Size. - Kaufman (1981) recently reviewed studies of LD subjects comparing the mean V-P discrepancies cited by the authors with the mean V-P discrepancies in the normal standardization sample. In the five articles he cited from the years 1976-1980 for which V-P discrepancies were available, the mean discrepancies ranged from 10.0 to 13.6 compared with 9.7 for the standardization sample. Kaufman concludes that "the V-P discrepancies for LD children have tended to be significantly (but not overwhelmingly) larger than normal values, although some studies have shown no difference at all . . . the data strongly imply that the magnitude of the V-P discrepancy . . . is not likely to be very useful in the diagnosis of LD or in its differential diagnosis." Other studies not included in Kaufman's review (e.g., Vance, Singer, Engin, 1980) further substantiate this conclusion.

There is, however, one group of LD students which seems to have a much larger discrepancy than normal, and this is the gifted LD. Schiff et al. (1981) studied a group of LD children who had either verbal or performance scores above 120. The mean fullscale IQ for the thirty children in this group was 123.0. The mean V-P discrepancy for the group was 18.6, or nearly twice the normal mean of 9.7. This was significantly larger than normal even for children with IQs above 120 (where the V-P discrepancy is 10.4 as cited by unpublished data in the Schiff paper). The authors acknowledge that there could be a selection bias toward a large V-P discrepancy as some clinicians give weight to the discrepancy in diagnosing LD, but this would also be true in average IQ LD populations which do not result in such large discrepancies. One possible explanation not elaborated upon by the authors would be that students would not have learning problems unless one or other of their subscale scores (either verbal or performance) was relatively low. That is, it seems likely that a person with a verbal IQ of 130 and a performance IQ of 115 would be less likely to have a learning problem than a student with a Verbal IQ of 130 and Performance IQ of 99. Setting the criteria of having one subscale score in the "gifted" range implies a large discrepancy if the other score is to be relatively low. More information about the sample and about normal high IQ children would be needed to sort out the full meaning of the large discrepancy size.

The question of discrepancy size and LD can also be approached by first identifying groups on the basis of WISC discrepancy and then looking for learning impairments in the groups. This approach was taken by Reed in 1967. In that study, WISCs were administered to all grade 1 (n = 248) and grade 5 (n = 233) children in three large public schools. At each grade level three groups were defined: VIQ$<$PIQ by 15 points or more (except at grade 1 where the criteria was 10 points because of the small number of subjects), PIQ$<$VIQ by 15 points or more, and V=P, in which the difference was no more than 2 points. An analysis was performed on reading scores with fullscale IQ covaried out. At age 6, there were no significant differences in reading scores. At age 10, the high verbal group had significantly higher (above average) reading achievement than the other two groups, which were not different. This suggests that the size of the discrepancy per se is not

associated with reading problems although the direction of the discrepancy may be related to certain achievements (as discussed in the next section).

In summary, when comparing LDs and normals, a large discrepancy cannot be considered either diagnostic or predictive of a learning disability. However, within certain groups of LD children there may be V-P discrepancy differences which are interesting and deserve further investigation.

Direction of the V-P Discrepancy. - Belmont and Birch (1966) did a review of studies of poor readers and reported that in all twelve studies for which V-P discrepancy data were available, verbal IQ was lower than performance IQ for poor readers. The authors, however, noted there were methodological problems with most of the studies including the fact that they were conducted on clinic populations. Belmont and Birch sought 9-10 year old subjects from a city in which standardized reading tests had been given. One hundred fifty readers were selected along with fifty control nonretarded readers. All of the children were administered a WISC, and it was found that 60% of the retarded readers had lower verbal than performance scores and 60% of the normal group showed the opposite pattern.

Very similar results were found by Huelsman (1970). He reviewed 23 studies for the period 1950-1970 and also found that about 60% of disabled readers showed a lower verbal than performance score. Huelsman then reports on his own study in which he reviewed the records of fourth grade children from several school districts and selected 101 with reading achievement below mental age level and 56 with reading achievement above mental age level. Among the disabled readers, 61% had lower verbal than performance scores. Among the reading achievers, 38% had lower verbal than performance scores.

Rourke and Finlayson (1978) assigned forty-five LD children to groups on the basis of their achievement not only in reading but also in math and spelling. The three groups (based on WRAT scores) were: (1) deficient in reading, spelling, and arithmetic; (2) deficient in reading and spelling but relatively good in arithmetic; and (3) deficient in math but adequate in reading and spelling. These groups were then compared on sixteen measures of auditory, verbal and visual-spatial abilities, among them being WISC Verbal and Performance scores. The findings included the fact that all subjects in group 3 (math deficient) had higher verbal than performance IQs. In group 2 (reading and spelling deficient), fourteen of the fifteen subjects had lower verbal than performance IQs. Group 1 (all deficient) also had all fifteen members with verbal lower than performance IQ.

Rourke's research group also did a series of studies looking first at the direction of the discrepancy in LD students and then determining which ability deficits were associated with those discrepancy patterns. The groups for each study consisted of high verbal-low performance (HV-LP), verbal equal to performance (V=P), and low verbal-high performance (LV-HP). The dependent variables included a receptive vocabulary test (PPVT), achievement tests (WRAT), and several subtests of the Halsted-Reitan. In the 9-14 year old group (Rourke, Young, and Flewelling, 1971), the scores of the HV-LP were superior to the LV-HP group on verbal, language, and auditory-perceptual skills. The HP-LV group was superior on visual-perceptual tasks, and the V=P group was roughly intermediate. Rourke and Telegdy (1971) used the same group criteria to measure complex motor and psychomotor abilities in LD 9-14 year

olds and found that the HP-LV group was superior on most measures. However, when both of these studies were replicated with a 5-8 year old group, the results were not so clear, and the authors argue for a guarded interpretation of the V-P discrepancies in younger children (Rourke, Dietrich, and Young, 1973). Wener and Templer (1976) objected to the statistical procedures of Rourke and Telegdy. When they replicated the study they did not find that most motor and psychomotor measures were superior in the HP-LV group, even at the 9-14 year old level.

Because low verbal scores have been most frequently associated with learning problems, a few studies have looked specifically at that group. Richman and Lindgren (1980) studied 81 children who had verbal deficits as identified by low verbal-high performance WISC patterns. All had been referred to an outpatient clinic, and the sample included all of those cases in which the VIQ was at least 15 points lower than the PIQ and the latter was at least 90. A factor analysis was done on the WISC, Hiskey, and WRAT subtest scores, and three separate groups were identified: A group with good abstract reasoning skills ($\underline{N}$ = 24), a group with good sequencing and memory skill ($\underline{N}$ = 19), and a group with general language deficits ($\underline{N}$ = 38). The general language group showed deficits in both abstract reasoning and memory. That group also showed poor reading achievement. The good abstract reasoners showed adequate reading achievement, and the sequencing-memory group appeared to be in between.
Richman and Lindgren's paper emphasizes some of the problems which have arisen in comparing groups formed on the basis of WISC discrepancy patterns. The groups do not appear homogeneous. Even within a group of low verbal clinical students, three separate groups can be identified which differ in ability and achievement. Similarly, as reviewed earlier, if the groups are formed on the basis of achievement, 60% of poor readers have lower verbal scores, but 40% do not. The direction of the V-P discrepancy is apparently too gross a measure.

Summary

One person can have a Fullscale IQ of 107 composed of an 87 Verbal IQ and a 130 Performance IQ. You might expect that person to learn and achieve differently than a person with the same Fullscale IQ composed of a 107 Verbal IQ and a 107 Performance IQ. This intuition has led clinicians and researchers in many fields, including learning disabilities, to look for the significance of the V-P discrepancy. As the above review indicates, however, firm conclusions still elude us. There is a tendency to expect and to find that low verbal scores are associated with auditory-verbal and reading problems, but the correlations are weak and inconsistent. As Rourke (1973) and others have pointed out, age of development is one variable that needs to be considered. Kaufman (1979) discusses a whole list of factors that may contribute to V-P discrepancy but are often ignored, such as the impact of a timed test, the issue of fluid-crystallized intelligence, and field dependence-independence. Finally, considerable attention has been given to the V-P discrepancy of normal achieving students. Rourke's studies of ability patterns in V-P discrepancies of the LD populations might profitably be applied to normal populations to see whether the same pattern of abilities prevails.

ASSESSMENT PHASE

Purpose and Overview

The assessment phase of the project was intended to provide three major types of information. First, by contrasting the performance of the LD and normal-achieving groups on measures of use of the isolation of variables strategy, the project provides data concerning the relative access on the part of the LD subjects to the more sophisticated problem solving approaches developed by normal-achieving subjects during early adolescence. The results of these analyses can be related directly to the discussions of previous researchers concerning cognitive development in this age range. A second type of information from the assessment phase relates to measures of task approach in addition to the propensity to control variables. The discriminant analyses based on these measures provide a more comprehensive picture of the approach taken to the tasks by the LD subjects and their normal-achieving peers than that provided by the analyses of unconfounded testing. The third type of information is related to the comparisons among the six subject groups. The inclusion of three LD subgroups and of control groups varying in age and IQ level provides an opportunity to address the issue of possible causes of the difficulties exhibited. In addition to data relevant to these three major issues, the present section contains data relevant to the issue of the generality of the findings from the bending rods task.

The organization of this section is as follows. The first subsection contains documentation of the need to study problems with thinking skills in the LD adolescent population. Data for this subsection were obtained from interviews with several LD resource teachers in the high schools used to collect the project data and from the questionnaires filled out by LD teachers to provide information concerning the major problem areas of the adolescents included in the LD sample. The second subsection contains a detailed presentation of the tasks and procedures used to collect the assessment phase data. Included here are descriptions of the subjects studied and procedures used to code various dimensions of performance in the rods and conductivity tasks. The third subsection contains a report of the major analyses used to contrast the performance of the rods task of the six groups in the study. These analyses included comparisons among the three LD subgroups as well as comparisons of each subgroup to normal-achieving subjects varying in age and IQ level. Results of both ANOVAs and discriminant function analyses are presented. Data relevant to the generality of the findings from the rods task are also included. The final subsection contains a description of the development and validation of a simplified version of the videotape coding procedures for use as a rating scale by LD teachers and clinicians.

Documentation of Need for Research on Thinking Skills

Teacher Interviews

What difficulties do high school LD students have with reasoning and problem solving in the classroom? To help answer that question from a teacher's perspective, eight high school LD resource teachers from five schools were interviewed. The first question of the interview was open-ended. The teachers were each asked to outline their primary concerns for the students they teach. If problem solving or reasoning was mentioned as a concern, the

following additional questions were asked: (1) What kinds of difficulties do your students have with organizational skills, reasoning, and problem solving? (2) What educational and behavioral approaches do you use to help your students overcome the effects of poor problem solving and reasoning? (3) What type of help would aid you in helping your students with these problems? and (4) Do you consider that poor reasoning and problem solving skills can be a primary component of learning disability, the secondary result of a learning disability, or are unrelated? If a teacher did not offer that reasoning or problem solving was a concern, he was asked whether it was. If the answer was affirmative, the same questions noted above were asked. The results of the interviews are summarized below.

All eight teachers indicated, either spontaneously or when asked, that problem solving and reasoning are areas of concern for their LD students. Of the fourteen "primary concerns" given spontaneously by the teachers, 43% related to educational vocational planning, 28% related to reasoning and problem solving, 14% related to emotional and social adjustment, and 14% related to the demands of parents or administrators.

When asked to describe the kinds of problems their students have with organizational skills, problem solving, and reasoning, teachers tended to think first of poor organization related to study skills: "Our students don't know what is assigned or when it is due." "They can't keep track of things--they come without pencils, papers, and books." "They don't seem to understand what the teacher wants." "They forget where they are supposed to go if the room is changed." Some teachers mentioned that even when students come with their materials and assignments, there are still difficulties. The student is apt to open to the correct page but not know how to get started or how to work independently. The students often don't know how to divide assignments into workable parts, how to find the main idea, or how to outline. They have difficulty following directions. As one teacher commented, "All our LD students are low in organization."

In contrast, the teachers seemed to feel the "reasoning" and "problem solving" are different from "organization" and that some students have difficulties with reasoning and problem solving but other students do not. One teacher said, "Low readers are the ones that have most difficulty with problem solving." Another teacher said that the students with behavior and learning problems were most apt to show poor reasoning "because they lack good judgment." Still another teacher felt that students from disorganized home situations are most likely to have reasoning difficulties. Several teachers named individual students who seemed to be excellent problem solvers in spite of severe learning disabilities.

When asked to describe the nature of the reasoning deficits shown by their students, the teachers mentioned difficulties with abstract thinking and causal reasoning in particular. Many teachers mentioned that abstract reasoning is required in high school subjects such as algebra and geometry and that these subjects may be especially difficult for some LD students. The teachers also emphasized that the reasoning deficits were not confined to academic areas. One student, for example, was observed to be so concrete in his thinking that he could not understand how a team labeled "the best" in a league could lose to a lesser team. Several teachers mentioned difficulties their students had with understanding the reasoning involved in driver

education. Problem-solving difficulties were also noted in finding the main idea, sorting out relevant from irrelevant material, and planning a problem attack strategy.

When asked how they help their students with difficulties in reasoning and problem solving, the teachers responded with techniques in two categories: (1) aids to organizing study and (2) methods to improve thinking skills. The first category was more highly emphasized and included such items as maintaining a class routine, providing structure, and making up study guides. Some LD teachers felt that they had to have frequent conferences with classroom teachers as a way of making sure that the students were following directions and conveying the teachers' expectations appropriately. Freshmen were felt to have special difficulties learning how to budget their time and set up a study schedule. Some students have to learn how to study. Teachers reported helping the students make out summaries, fill-in the blank self-test questions, and outlines. The goal of all of these activities, from the teachers' standpoint, was to help the students "learn how to learn" at the high school level.

The second category, teaching thinking skills, included efforts by the teachers to help the students do more than memorize facts. One teacher stated that she tries to get her students to focus on how they go about solving problems. She has them talk through a problem with her and then apply the method to a new problem. Other teachers mentioned using puzzles or matrices to help the students think logically. Science activities which employ manipulative materials were also thought to be helpful for some students. Another teacher stated that he likes to use analogies, showing step by step relationships between things as a way of leading students to logical reasoning. Consumer education classes were mentioned as one of the ways in which practical reasoning is taught. There students work on real life situations such as how to make a good decision in buying a car. One teacher mentioned that his class also writes "Dear Babby" letters which are practical problem solving essays. They emphasize how to make good decisions and wise choices.

When asked how successful they were at teaching reasoning and problem solving skills, the teachers expressed some uncertainty. They all felt that they needed more time and fewer students to be really effective. Teachers at some schools complained that they had to focus on content and on each day's homework and that therefore there was no time to work on thinking skills. Other teachers suggested that more homogeneous groups (i.e., homogeneous with regard to reasoning and problem solving) would be helpful. Putting a priority on these skills at the high school level was also suggested. Finally, the teachers noted that there was no standardized way to measure thinking skills or progress in reasoning. For this reason, it tends to be overlooked on evaluations and I.E.P.s.

When asked whether they thought problems with reasoning were an effect of a learning disability or a learning disability in themselves, the teachers gave varied responses. One teacher mentioned that these problems do not only effect LD students. High school is less structured, and the content of the courses is more abstract. Therefore many students who are not LD begin to have problems. Other teachers pointed out that it is a matter of definition

and that at some point problems with thinking have to be called learning disabilities because they impair learning.

In summary, it is clear from the comments of these eight LD teachers that they had all thought about reasoning and problem solving as areas of deficit in some of their LD students and that they felt that these areas should concern them as LD teachers. To the extent that time allowed, they had each devised ways of helping their students remediate deficits in thinking skills. However, in spite of the fact that all of the teachers could readily come up with anecdotes to illustrate their students' poor reasoning skills, there was great diversity among the teachers in priorities. As one teacher put it, "learning to reason well is one of the most important survival tools that students can learn." That teacher spent a high proportion of her time on reasoning skills. Other teachers felt that their school districts did not recognize the importance of remediating thinking skills. Those teachers felt they had to spend their time primarily on homework and organizational techniques. Despite their own position on this issue, all teachers interviewed seemed enthusiastic about research which would help define the kinds of reasoning problems that they observe more informally in the classroom setting.

Data from Teacher Questionnaires

The questionnaires completed for twenty of the twenty-seven learning disabled students in the study reveal that LD teachers' concerns about reasoning and problem-solving deficits in the (LD) population as a whole are applicable at the individual level as well. When asked to list each student's major difficulties in order of severity, the LD teachers rated 8 out of 20 students (40%) as exhibiting some aspect of thinking, reasoning or organizational skills as a primary difficulty. When asked specifically about problems in thinking skills, 12 out of 20 students, or 60%, were felt to have thinking skill deficits. Two additional subjects were described as demonstrating deficiencies in specific subskills subsumed by the area of thinking skills. Information about the kind of thinking skill problems most frequently noted by the LD teachers was also obtained. When asked to select phrases that could be used to describe their students' problems in this area, the teachers chose "difficulty judging relevant vs. irrelevant information" most often. Sixty-five percent of the students were described as demonstrating mild to severe difficulty here. "Difficulty with logical reasoning" and "study skill problems" were also frequently noted. Fifty-five percent of the students were felt to have some degree of difficulty in these areas.

General Procedures

Subjects

Subject Group Criteria. - The goal of subject selection was to find ten subjects for each of three learning disability groups and three control groups. The groups were designated as follows:

1. No discrepancy LD (ND)
2. Low verbal LD (LV)
3. Low performance LD (LP)
4. High IQ ninth grade control (High)
5. Low IQ ninth grade control (Low) } (N9)
6. Fourth grade control (N4).

Groups 1-5 were all high school freshmen, ages 14-15 years. Group 6 consisted of fourth graders, ages 9-10 years. All subjects in the three learning disability groups had to have been identified by the school as learning disabled and had to be receiving LD services at the time of this study. All control subjects had to be students who had not been identified as learning disabled and who were not receiving special education services.

Scores on the Wechsler Intelligence Scale for Children-Revised (WISC-R) were used as the other major component for selection into groups. The specific WISC-R criteria for each subject group are listed below.

Group 1 - No Discrepancy LD
a. WISC-R fullscale IQ score 85-115
b. Difference between VIQ and PIQ of 9 points or less

Group 2 - Low Verbal LD
a. WISC-R fullscale IQ score 85-115
b. PIQ > VIQ by at least 15 points
c. PIQ score 85-120
d. Vocabulary or comprehension subtest score 8 or lower
e. Block design & object assembly subtest score 9 or higher

Group 3 - Low Performance LD
a. WISC-R fullscale IQ score 85-115
b. VIQ > PIQ by at least 15 points
c. VIQ score 85-120
d. Block design or object assembly subtest score 8 or lower
e. Vocabulary and comprehension score 9 or higher

Group 4 - High IQ Ninth Grade Control
a. WISC-R fullscale IQ score 100-120
b. VIQ = PIQ (9 or fewer points discrepancy)

Group 5 - Low IQ Ninth Grade Control
a. WISC-R fullscale IQ score 80-90
b. VIQ = PIQ (9 or fewer points discrepancy)

Group 6 - Fourth grade control
a. WISC-R fullscale IQ score 85-115
b. VIQ = PIQ (9 or fewer points discrepancy)

Recruitment. - Subjects were recruited from suburban schools in the Chicago area. The general process required several steps: (1) A school district was contacted and permission was requested to conduct the study; (2) If permission was granted, the study was discussed with relevant school staff members (usually LD coordinators, teachers, and/or counselors). (3) A pool of students was identified as possibly fitting the criteria of either control or learning disabled subjects. (4) The study was described to these students in writing (and also orally in most cases), and the students and parents were asked to give written consent if the student was willing to participate. (5) After permission had been granted, testing of the subject was scheduled. All participation was voluntary and could be terminated by the students or parents at any time.

Because it was not possible to obtain all the subjects needed for each group of this study at one school, additional school districts were contacted until subject selection was as complete as possible under the time constraints. A total of nine high schools and four elementary schools provided the subjects. Six additional high school districts had been contacted but did not consent to participate.

Some adjustments were made in the recruitment procedure over the months of testing to help balance the groups. For example, when it became evident that many more boys than girls would be participating in the LD group, but that girls were more apt to volunteer for the control group, recruitment of controls was eventually limited to just boys. When it became evident that most of the normal students who agreed to participate had higher level IQs, attempts were made through the school staff to recruit students who were likely to have lower IQs.

In spite of the recruitment efforts and cooperation of school systems, we fell short of our goal for subjects in two areas: (1) low performance LD students and (2) low IQ ninth grade controls. In all of the high schools, LD coordinators indicated that low performance LD students were underrepresented in their classes because most did not show significant deficits in academic areas such as reading and therefore were not served by the LD program. The lower IQ ninth grade controls were also hard to find, probably because they were less willing to volunteer. Also, those who did agree to participate seemed more likely to miss appointments.

Of LD students who were originally contacted, approximately 50-75% agreed to participate in the study. The rate of volunteers was highly dependent on interest and influence of the students' LD teachers. Many teachers strongly encouraged their students and reminded them of their appointments. In some of those cases, participation was very high.

The control subjects were in many respects much more difficult to recruit. Only 20-30% of ninth grade students originally contacted agreed to volunteer or remembered to come to their appointments. Of those who agreed to participate and who were tested on the WISC-R, about 50% had a discrepancy between verbal and performance IQ that was greater than 10 points and were not eligible for the study. Even among the fourth graders, where principals and teachers were asked to help identify children who were apt to fall in the national average for IQ, only 40% of those tested were within the appropriate range and had 10 points or less discrepancy.

Characteristics of Final Subject Groups. - Table 1 shows the composition of each group by subject. Table 2 presents a summary of group compositions. For each learning disabled subject included in the study, a school LD teacher was asked to complete a questionnaire rating the student in several academic areas (see Appendix). The purpose of the questionnaire was to help clarify the kinds of learning problems demonstrated by each of the study's LD groups. Eligibility for each group had originally been determined on the basis of each student's scores on the Wechsler Intelligence Scale for Children-Revised (WISC-R). It was assumed that students who had low nonverbal scores relative to their verbal on the WISC-R would probably show nonverbal learning problems in the classroom. It was also assumed that students who showed the opposite WISC-R pattern (i.e., low verbal relative to nonverbal) would show language

problems in the classroom. Those LD students who had no discrepancy between verbal and nonverbal WISC-R scores were assumed to have a more circumscribed decoding problem rather than either a general language or nonverbal deficit. The purpose of the teacher questionnaire was, first, to provide an independent assessment of the extent to which these assumptions would be supported by classroom behavior. Secondly, questions about thinking skills were asked in the questionnaire to see whether teachers perceived one or other of the groups to have more difficulty with some aspect of this general area.

The questionnaire (Appendix) first asked the teacher to list the student's major learning problems in order of severity. The teacher was then asked specifically whether the student had problems in each of five areas: language, reading, nonverbal, math, and thinking skills. In each area the teacher was asked to describe the nature of the problems, if any, and to indicate the severity (mild, moderate, or severe) of impairment in a series of subskill areas.

The questionnaires were handed out to teachers of all twenty-seven learning disabled students included in the study. If the questionnaire was not returned, a second questionnaire was mailed to the teacher with a reminder. Using this method, twenty questionnaires were returned for the twenty-seven students. That is a return rate of 74%. Nine questionnaires were returned from the ten students in the no discrepancy group, five out of ten were returned from the low verbal group, and six of seven were returned from the low performance group.

Table 3 summarizes the teachers' responses when asked to state their student's most serious difficulty. In the no discrepancy group, six of the nine questionnaires (67%) listed some aspect of thinking, reasoning, or organization as the student's primary difficulty. In each of the other two groups only one student (17-20%) was listed as having a thinking skill deficit as his/her major problem. In the low verbal group, reading and syntax were listed as major problems and in the low performance group, motor, visual-spatial, and attention skills were listed as the weakest areas.

Table 4 summarizes the teachers' responses when they were asked whether or not the student had a problem in each of the following areas: language, reading, nonverbal, math, and thinking skills. Note that both the no discrepancy and low verbal groups have less than 50% showing nonverbal problems. In contrast, in the low performance group 100% of the members are characterized by the teachers as showing nonverbal problems, while 33% or fewer members have either a language or reading problem. All groups show 60% or more of their students with math problems. When asked specifically about problems in thinking skills, the no discrepancy group is thought by the teachers to have most difficulty (which is consistent with the data reported in Table 3). Here 100% of the students were rated as having thinking skill problems. In the low verbal group only 40% of the students were thought to have thinking skill deficits, and 50% of the low performance group show those deficits.

Table 5 lists specific subskills included in each academic area to further clarify the nature of the students' problems. Considering those subskills in which at least 65% of the group are rated as moderately or severely impaired, the no discrepancy group shows greatest deficits in comprehension (both language and reading), math word problems, and organizational and study

skills. The low verbal group also shows a high level of impairment of language and reading comprehension but decoding is added as an area of major deficit. Only 40% of the low verbal group shows organizational or study skill problems as compared to 78% of the no discrepancy group.

Table 5, consistent with Tables 3 and 4, shows that neither the no discrepancy group nor the low verbal group shows large percentages of students with any problem in nonverbal areas. By contrast, in the low performance group more than 65% of the students show significant impairment in three nonverbal areas: visual-motor integration, orientation, and social perception. Like the no discrepancy group, the low nonverbals also show impairment on math word problems and organizational and study skills.

Because of the limited number of questionnaires returned by the teachers, particularly in the low verbal group, the results of this analysis have to be considered as suggestive rather than definitive. However, it seems clear that the group that was defined by the WISC-R as the "low performance" group, shows ample evidence to the teachers of having nonverbal learning problems, while neither of the other groups shows those kinds of problems. Distinguishing between the academic disabilities of the other two groups is more problematic. Both groups are rated as having significant reading deficits and some language problems. The teachers found more problems with decoding in the low verbal group than in the no discrepancy group. The opposite had been assumed when the groups were originally composed. However, the groups are further distinguished by the fact that teachers listed reading and language problems as the greatest area of deficit for most of the low verbal students but not for most of the no discrepancy group. Therefore, the assumption of classroom verbal problems in the group that was originally composed as "low verbal" on the basis of the WISC-R seems to be supported. However, the learning problems of the no discrepancy group are less clearly delineated by the teacher questionnaire. The teachers named reasoning and organizational problems as the most severe deficit of this group, a deficit that was shared by 100% of the members. Perhaps that is the best characterization of the group.

Ranking the three groups in terms of thinking skills, the teachers' ratings indicate that the no discrepancy group is the most severely impaired, the low performance group is next, and the low verbal group is the least severely impaired. In both the no discrepancy and low performance groups, the kind of thinking problem most frequently noted was organizational and study skill deficits.

Task Materials and Administration Procedures

Schedule Overview. - All testing was done individually at the school during a student's free period or LD resource period. Each session was forty to fifty minutes long. The first session was used to administer the WISC-R to all control subjects and to any LD subject who did not have a recent (within three years) WISC-R on file. Another session was used to administer the rods experimental procedure, and the final session was used to administer the conductivity experimental procedure and the Concept Formation Task of the Woodcock-Johnson. Both the rods and the conductivity sessions were videotaped. No specified time interval occurred between the WISC-R and rods sessions, but the conductivity session was scheduled between one and two weeks after the rods session. If the student or his family wished to view the

videotapes or discuss the student's performance, this was arranged after the final session.

Wechsler Intelligence Scale for Children-Revised. - The WISC-R was administered in the standard manner by a psychologist or graduate student who was proficient in its use. Where time permitted, all ten of the regular subtests were administered. For eight subjects where time was limited, eight subtests were administered (eliminating Similarities and Picture Arrangement), and prorated scores were obtained. The WISC-R was scored by two raters, and discrepancies in scoring were resolved by discussion.

Bending Rods Task. - Materials for the rods task consisted of 26 rods (two sets of 12 rods each and one additional pair of rods used during the initial task instructions), 2 pairs of weights to attach to the rods, and a stand into which the rods could be inserted. The rods were selected to vary along five dimensions (i.e., variables) with two or three levels of each dimension: length (20, 40, and 60 cm), diameter (.32, .28, and .64 cm), material (plastic, wood, and steel), base (black plastic or metal). Amount of weight (metal washers of two different sizes) also varied. The rods and weights were identical on all other possible dimensions. The two sets of rods differed only in terms of the particular combinations of levels of the dimensions represented. In both sets each level of each dimension occurred at least once. Also, in each set there were the same number of possible unconfounded tests of each variable and approximately the same number of possible confounded tests. Thus, the two sets did not differ appreciably in the probability with which a subject could select at random a pair of rods which would constitute an unconfounded test of any given variable.

All subjects were given the following sequence of subtasks in the same order: Test Construction 1, Test Choice, Test Construction 2, and Test Evaluation.

In the Test Construction Task, the subject was presented with a set of 12 rods, the four weights and the stand. The pair of demonstration rods, which constituted an unconfounded test for length, was used to demonstrate that some rods bend more than others. The subject was then asked to figure out, by comparing pairs of rods, what makes some rods bend more than others and to explain why he had chosen each pair of rods and what he had learned after each test. Follow-up questions were used as needed to clarify the subject's choice of rods. After the subject indicated that he was finished testing rods on his own, the experimenter asked for a summary of conclusions about bending. If fewer than five dimensions were mentioned by the subject, he was told that some other things matter for bending and asked what they might be. Any dimension not identified by the subject spontaneously or in response to this question was then named by the experimenter and the subject was asked to show how that dimension influences bending. Further tests of the effect of some of the dimensions were then requested by the examiner, if the subject had not made consistently confounded or unconfounded tests of that dimension prior to the summary.

In the Test Choice Task, the subject was presented sequentially with five sets of two pairs of rods each. Each set offered both a confounded and an unconfounded test for one of the five variables. Here not only were the rods preselected and the variable to be tested specified, but also the comparison of two pairs was likely to call attention to the confounding variable. For

each pair-set the subject was asked, "Which set do you think is better for finding out if _____ matters? Why is that set better?" The subject was asked to respond without using the weights except for the test of weight, when the experimenter placed the weights on the rods.

In the Test Evaluation Task, the subject was presented sequentially with six pairs of rods (only one of which constituted an unconfounded test of a variable). For each pair, the subject was asked, "Could you use these rods to tell for sure whether _____ makes a difference? Why or why not?"

Chemical Conductivity Task. - Materials for the conductivity task consisted of 14 glass vials of chemical solutions sealed with black rubber stoppers and a lucite box for conducting tests. Extending through each stopper were two metal electrodes which extended into the solution. The top of each electrode was curved to facilitate the attachment of an alligator clip. A subject could test two vials at a time by placing them in the box, attaching the electrodes of each vial to a set of wires with alligator clips, closing the lid of the box and pressing a switch. The amount of conductivity through the solutions could then be assessed by comparing the relative brightness levels of two light bulbs on the top of the box.

All subjects were given a task similar to the Test Construction 1 subtask of the rods. No other subtasks were administered.

Woodcock-Johnson Concept Formation Test. - This was administered and scored according to the Manual's directions. (Woodcock & Johnson, 1977) Standard scores (based on age expectations) were used in subsequent analyses.

Other Data. - Information gathered from the school files included the following: birthdate, grade, parental occupation, recent achievement scores, WISC-R scores (when available from previous evaluations), and a description of the learning disability. In addition, the teacher of each LD student was asked to complete a questionnaire about the student's academic skills. Information was filed by code number to protect the students' privacy.

Data Coding

Overview of Coding. - The data from each experimental session consisted of (1) a protocol form on which the examiner had kept a record of materials used by the subject (i.e., which rods, weights, and bottles) and had made preliminary notations about the performance and (2) videotape of the session. The videotape was returned to the laboratory and copied. The copying procedure included the addition to the tape of a continuous time marker. An audiotape of the session was also made from the videotape. A transcript was typed from the audiotape and then edited by viewing the videotape. Time markers were added to the transcript to facilitate matching videotape and transcript.

Coding was done by using the protocol, transcript, and/or videotape depending on the requirement of the particular code. The sections which follow provide a brief description of the coding procedures used to obtain the dependent variables used in the major analyses. This description is intended only to provide a general sense of the factors which entered into the coding decisions. For more details concerning specific coding criteria as well as examples of coded utterances, the reader is referred to the Coding Manual.

Preliminary Coding. - This stage of the coding process involved decisions about which variables the subject mentioned during each episode and which of those variables mentioned were the object of the test being conducted. In order to reach these decisions, two types of information were used: a description of the rods chosen for the test, and the verbal expressions used by the subject to refer to the rods and to the outcome observed. A series of criteria were developed which allowed coders to make decisions concerning the clarity and referent of each expression.

The verbal expresions for each test were first divided into three categories:

1) variable expressions - those which were clearly references to the rods or to rod attributes [example: the metal one];
2) outcome expressions - those which were clearly references to the differential bending of the rods [example: is bending more];
3) bridging expressions - those which could be interpreted as either variable or outcome expressions [example: is more flexible].

Expressions in these three categories were recorded in three separate columns of a coding form. All expressions accompanying a given test were recorded in a single section. Within that section, those expressions relevant to a single variable were recorded on the same row. Thus, there was one row in each section for each variable mentioned during that test. Outcome expressions were recorded in the row corresponding to the variable with which they seemed to be paired. (See the Coding Manual for a discussions of the treatment of the bridging expressions.)

When all of the referring expressions accompanying a given rod pair had been recorded, a decision was made concerning which variable or variables was the object of the test (if any) being conducted by the subject during that episode. This decision was based primarily on the pairing of variable and outcome referring expressions. Finally, for each rod pair, a decision was made concerning whether the test was "confounded" (i.e., more than one dimension was varied) or "unconfounded" (i.e., only one dimension was varied). This decision was based on the rods chosen and, in the case of unconfounded tests, on the subject's conclusions about his test.

Measures of Preference for Unconfounded Tests. - Four indices of a subject's preference for unconfounded testing were coded. Each subject's proportion of unconfounded tests was computed separately for TC1 and TC2. These proportions constituted the first two indices of unconfounded testing.

The remaining two indices of unconfounded testing consisted of the number of correctly justified judgments for the Test Choice and Test Evaluation phases. Each justification was scored on the basis of whether or not the subject explained the necessity to control variables or objected to a test because the rods were confounded.

Additional Measures of Task Approach. - Six additional estimates of the subjects' general task approach were examined. Five of these measures concerned the subjects' verbal reasoning strategies and one assessed their response to examiner guidance. These measures were applied to each subject's performance during each Test Construction phase.

The five reasoning variables were derived from ratings made of the subjects' predictions and conclusions during each of their spontaneous tests. Each of the original rating scales consisted of three, four, or five ordinal categories. These rating scales were converted to proportion scores and the "best" proportion variable from each of the five reasoning scales was selected. These "best" proportion variables were those that showed a reasonable degree of variance, were not highly intercorrelated with other variables, and exhibited a relatively high degree of interobserver reliability (intraclass correlations at or above .80). The resulting five variables are described below.

1. The proportion of predictions that exhibited a high degree of intentionality. Predictions were statements made by the subject in response to the examiner's question, "Why did you pick these?" High intentionality was defined as making either empirically based predictions or hypothetical predictions.

In empirically based predictions, differing rod attributes are mentioned and reference is made to the need to note relative outcome. In addition, a strong sense of intent in conveyed through the use of certain key phrases such as "I wanted to compare," I was seeing if," or "I wanted to see." Although an a priori prediction regarding bending is made, the prediction only relates to the specific rods at hand. Such predictions are not hypothetical in the sense that they do not infer anything about the relative bending of rods in general.

Example: "I was seeing if this one was the metal one and this a wood one. See which one pulled out longer."

As with empirically based predictions, in hypothetical predictions, rod or bottle attributes are mentioned and reference is made of the need to note relative outcome. A sense of intention is conveyed through the use of the key intentional phrases. Unlike empirical predictions, however, hypothetical predictions are concerned with whether or not a given variable makes a difference for the outcome in general, that is, for the set of all possible rods or bottles, not only for the two in the stand at that time.

Example: "I picked these because they're the same length, different material, same base. So I wanted to see if the material would make a difference."

2. The proportion of predictions that exhibited a low degree of explicitness. Low explicitness meant that no variable was specified in the prediction.

Example: "I wanted to see which one would bend more. They're both wood and they both look like the same length." [Subject does not mention which variables differ.]

3. The proportion of conclusions that included a highly articulated description of both rod variables and bending. The purpose of the articulation code is to describe the degree of conceptual sophistication of the relationship between rod attributes and bending. Each category rates the level of detail of both attributes and bending and the nature of the relationship between the two. This code is seen as an important piece of the data which subjects use to make their causal inferences. At the highest

level, the subject conceptualizes the relationship as a continuous one as revealed by his use of inclusive terms which capture all the levels of an attribute and all degrees of bending.

This category includes the following examples:

a. When both ends of the dimension are mentioned;

e.g., "The thick one bends less than the thin one."

b. When both the attribute and the bending expression are expressed in relative terms.

e.g., "The thicker one bends more."
"Thinner ones bend a lot."
"The wood bends down far."
"The steel one stays the same."

c. When every possible level within a variable is exhausted by the subject's referring expression, and bending is not only expressed in relative terms, but in a way which includes all states of bending.

e.g., "Thickness matters for bending."
"The longer they are the more they bend."
"The less weight on it, the less it goes down."
"Steel bends the most, then wood, then plastic."

d. When a relationship exhausts all possibilities within a specified level of another variable.

e.g., "The longer the plastic, the more flexible it is."

4. The proportion of conclusions in which no attempt to derive an explanatory principle for differential bending was made. Explanatory principles consist of attempts to account for differential bending by referring to unseen constructs such as density.

Examples of conclusions in which explanatory principles are absent:

"Long ones bend more."
"This one I knew was gonna bend and this one, I knew wasn't gonnna bend."
"The thinner the plastic is, the more it will bend."

5. The proportion of conclusions that exhibited a low degree of generality. Subjects received a low score on generality if their conclusions referred to specific rod pairs only. Such verbal specifics as this rod or the long one or use of the present progressive usually indicate a specific reference.

Examples: "This long one bends more."
"The plastic is bending more and with an arc to it than the metal."
"Different bases. This one bent a little bit."
"A lot more weight on it and it goes down."

6. The proportion of examiner-requested tests that resulted in noncompliance. For example, suppose that the examiner asked for a test of length. If the subject selected two rods equal in length but of different diameters, and drew a conclusion only about diameter, he or she would be given a noncompliance score for this test.

The above six measures provide estimates of subject's higher-level verbal expressive and receptive skills. The first five measures code the degree of sophistication seen in their verbal predictions and conclusions. The last measure indicates how well they understand the examiner's test of a particular variable, such as length.

Observer Agreement and Reliability Estimates. - Observer agreement figures and interobserver reliability estimates were calculated on all the codes described in this report and on several other codes which were implemented but not analyzed due to low reliability or poor subject-variable ratios. These calculations were done in two steps. First, observer agreement with a criterion or expert coder was computed for the coding of the preliminary coding forms. Second, interobserver reliability estimates were computed on all other codes.

This two-step procedure resembles that recommended by Frick & Semmel (1978) for observer training and reliability estimation. The first step, observer agreement assessment, enables one to estimate the ability of individual coders to agree with an expert coder on the coding of the primary coding form, the referring expression sheet. This form contained data used by several other codes and thus its accuracy influences the accuracy of all subsequent codes. After the referring expression sheet agreement statistics were examined, it was decided that accuracy would be increased if all these sheets were coded independently by two observers. Then, these observers would resolve any differences consentually. Therefore, this primary coding form reflected the collective judgments of two observers.

The second step in the procedure yielded the interobserver reliability estimates for the codes used in analyzing the data. Intraclass correlations were computed on the same coding units that constituted the variables used in the data analysis (Frick & Semmel, 1978; Johnson & Bolstad, 1973).

Finally, these estimates of observer agreement and interobserver reliability were made prior to and during actual data coding. This was done to minimize the possibility of a decrease in observer accuracy between the end of training and the beginning of data collection (Taplin and Reid, 1973). In addition, observers were unaware which of their observations would be used to assess agreement or reliability. Each observer coded a sample of videotapes independently and then a subset of these same tapes was randomly selected for coding by another observer. This procedure controlled for the possibility that knowledge of reliability assessment would temporarily increase observer accuracy (Reid, 1970; Romanczyk, Kent, Diament and O'Leary, 1973; Taplin and Reid, 1973).

Four observers were involved in coding the rods preliminary coding forms. A subset of all the rods sessions was coded independently by each of the observers. As the initial step in establishing observer agreement, one of the four coders was designated as the criterion or expert observer. Estimates of

agreement between each of the remaining three coders and the criterion were then obtained by the following procedure. Each pair of coders (the criterion and one other) randomly selected five cases from among those that one member of the pair had previously coded. Either test construction phase one or two of these five cases was then independently coded a second time. A measure of observer agreement for each coder with the criterion was obtained by tallying the number of agreements concerning the variables recorded as being mentioned (e.g., length, material, diameter, weight, base and nonstandard variables) and tested in each of the five cases. Percent agreement for each case was calculated by dividing the number of agreements by the total number of agreements plus disagreements. Percent agreement across all five cases was then averaged to yield a single index of agreement for each observer with the criterion. Agreement statistics are shown in Table 6.

As can be seen, average percent agreement ranges from .75 to .79. Although these figures are somewhat low, they represent agreement per case on nonstandard and ambiguous variables as well as on the standard variables (see Coding Manual) that were used in data analysis. Since many disagreements involved nonstandard or ambiguous variables rather than standard variables, agreement rate for the variables included in the analysis was higher than reported.

Because the preliminary coding sheet provided the data base for most of the other codes in this study, it was important that this sheet reflect our best judgment of a child's performance. It was therefore decided that each session should be coded independently by two observers first. All disagreements that resulted from the double coding were discussed and resolved. The final preliminary coding sheet on which all other codes were based was therefore coded consentually.

Two coders were involved in coding the conductivity preliminary coding form. To establish observer agreement a subset of cases was independently coded by one coder. Then the other coder randomly selected five cases from among those previously coded. A measure of observer agreement was obtained by tallying the number of agreements concerning the variables recorded as being mentioned and tested in each of these five cases. Percent agreement for each case was calculated by dividing the number of agreements by the total number of agreements plus disagreements. Percent agreement across all five cases was thus averaged to yield an overall index of agreement. Overall agreement per case averaged .80 and ranged from .72 to .89. Agreement figures by case are shown in Table 7. The conductivity preliminary coding sheets were all double coded and disagreements were resolved by consensus.

The summary codes of unconfounded testing obtained from both the rods and conductivity preliminary coding forms were: the total number of tests, the total number of unconfounded tests, the number of unconfounded tests before the summary, the number of unconfounded tests in response to probes, the total number of unconfounded variables, the number of uncredited unconfounded rod pairs, and the number of tests with counteracting variables. Intraclass correlations were computed for each of these summary codes. Intraclass correlation is the suggested method of establishing reliability for interval (Kirk, 1968; Winer, 1971). The intraclass correlations for the summary are shown in Tables 8 and 9. Only two coders were involved in coding nonverbal summary codes (i.e., number of uncredited unconfounded tests and

the number of tests with counteracting variables), hence only one reliability estimate was computed for each of those codes. As can be seen, the intraclass correlations of these summary codes were between .75 and .99 (mean = .85).

Two observers were involved in applying the rod scanning and selecting code. To obtain a measure of interobserver agreement a subset of five cases was randomly selected from cases that had been coded by one of the two observers. Each of the five cases was then coded again by the observer who had not originally coded the case. Intraclass correlations were then computed on the variables which comprise the scanning code. The resulting correlation coefficients are shown in Table 10. There was a wide range of reliability on these variables. Variables with low reliability (i.e., below .75) were not included in subsequent data analyses.

The remaining codes implemented were articulation, generality, explanatory principles, and prediction (reasoning codes) and rod attributes (data gathering code). These codes were recoded to change their level of measurement (from ordinal or nominal to ratio). Each of these codes was recoded as a series of proportion scores.

For example, in the generality code, each conclusion made about bending may receive one of three ratings (from low to high generality). The generality summary codes used in the data analysis were the proportion of each subject's tests which received a rating of one, the proportion of a subject's tests which received a rating of two, and so on. This procedure was used so that ratio scores could be computed from ordinal scales.

For each of the reasoning codes, a single coder coded all cases. In order to establish interobserver reliability for these codes, another coder learned how to apply the code by studying the procedures in the coding manual. This second coder then randomly selected five cases that the initial coder had previously coded and independently coded them. Reliability estimates for each of the reasoning codes were derived by computing intraclass correlations on the proportional summary scores. The correlation coefficients for the summary codes are shown in Tables 11 - 16. For the most part, the intraclass correlations on these proportion scores were high (between .88 and 1.00). Some of the very low correlations were obtained for very infrequent codes (e.g., a score of "2" for Generality). Summary codes having low reliability (below .80) were not included in subsequent data analyses.

Group Differences in Use of the IV Strategy

Overview of Analyses and Technical Considerations

The analyses to be reported in this section were intended to provide information relevant to the general question of group differences in the use of the IV strategy (see Objectives 1 and 2). Several possible indices of use of the IV strategy were available for this analysis. The index used here was the proportion of all spontaneous and probed tests which were unconfounded. This measure was chosen because (a) it reflects the subject's testing of all of the variables (those spontaneously identified and those pointed out by the examiner) and (b) it provides a wider range of scores than would an index of the number of variables (out of five) tested in an unconfounded manner (used in some previous studies, e.g., Day and Stone, 1982).

One approach to answering the question of differential IV strategy use would involve a single analysis of variance which contrasted the performance of all six groups in the study. Although appealing because of its simplicity, this approach was not adopted. Since the focus of the project was on contrasts among specific groups rather than on overall group differences (see Objectives), such an analysis would have been superfluous. Also, planned comparisons of relevant groups would not have been possible because of non-orthogonality. Thus, a series of seven smaller ANOVAs was used to analyze the data. The first ANOVA provided information concerning diferences among the three LD subgroups. The remaining six ANOVAs fell into two clusters. Within each cluster of three ANOVAs, each LD subgroup was compared in turn to appropriate control groups. In ANOVAs 2-4, the LD groups were compared to same age and younger peers matched for fullscale IQ. In ANOVAs 5-7, the LD groups were compared to same-age peers of two IQ levels.

All seven analyses were conceived as 3 X 2 mixed ANOVAs, with group membership as the between-subjects factor and test-choice administration (TC1 vs. TC2) as the within-subjects factor. The actual analysis proceeded in two stages, with a one-way ANOVA on TC1 scores, followed by a second one-way ANOVA on TC2 scores, using TC1 scores as a covariate. This design allowed the evaluation of group differences in initial use of the IV strategy in the rods task and the evaluation of differential improvement during the session, controlling for initial level of performance. A detailed description of each of the seven analyses is provided below. Following the detailed presentation of the results is a summary of the general findings. Readers interested in a general overview of the results should move directly to the summary section.

Group Differences in Task Success

Comparisons Among Three LD Subgroups. - The analysis of overall group differences in the tendency to construct tests in an unconfounded manner during the two test construction phases was conducted using a series of ANOVAs and ANCOVAs with the proportion of unconfounded tests as the dependent variable. Planned comparisons of differences among means were conducted using Dunn's procedure with the experiment error rate set at .05 (Kirk, 1968). The results of these analyses are summarized in Tables 17 and 18.

The ANOVAs on performance during the first and second test construction phases (Table 17) revealed no significant differences among the groups in the tendency to construct unconfounded tests, F (2,24) = 1.12, $p > .05$ and F (2,24) = 1.12, $p > .05$, respectively. However, as can be seen in Table 18, there was a tendency for the ND group to construct fewer unconfounded tests than either of the high discrepancy groups. Effect sizes (Table 20) indicate that the differences between the ND group and each of the other groups are moderate to large (Cohen, 1977), while the difference between the two high discrepancy groups is considerably smaller.

A test of differential improvement from the first to the second test construction phase was conducted using an analysis of covariance on final performance (TC2) with initial performance (TC1) as the covariate (see Table 17). Results revealed significant improvement across test phases, F (1,23) = 22.47, $p < .05$, but no significant differential improvement among groups, F (2,23) = .53, $p > .05$.

Analyses of the more structured test choice and test evaluation phases of the rods task produced information concerning group differences which was different from that reported above (see Table 19). There were significant group differences for the test evaluation questions, F (2,24) = 3.72, $p < .05$. The LP group performed significantly better than the LV group. Note also that, contrary to the results from other measures, the LV group scored lower than the ND group. While there were no significant group differences on the test choice questions, F (2,24) = 1.52, $p > .05$, effect sizes (Table 20) indicates that the performance of the LP group is high relative to the other two groups. Note that this pattern of group differences across the four phases of the rods task indicate that the ND group is at a disadvantage during test construction, and the LV group at a disadvantage during the more structured question phases.

Comparisons With Same Age and Younger Controls: ND vs. N9 vs. N4. - The analysis of overall group differences in the tendency to construct tests in an unconfounded manner during the two test construction phases was conducted using a series of ANOVAs and ANCOVAs with the proportion of unconfounded tests as the dependent variable. Since one would expect overall significant effects to appear by virtue of the inclusion of fourth and ninth normal-achieving subjects in the design, planned comparisons of the crucial differences involving LD subjects were conducted using Dunn's procedure with the experiment error rate set at .05 (Kirk, 1968). The results of these analyses are summarized in Tables 21 and 22.

The ANOVA on performance during the first test construction phase (TC1) indicated significant differences among the groups, $F(2,34) = 3.56$, $p < .05$. As can be seen in Table 22, there was a tendency for the ninth grade control subjects (N9) to conduct a higher percentage of unconfounded tests (38%) than either the LD subjects (22%) or the fourth grade controls (N4) (20%), whose performance was similar. Pair-wise tests of these differences fell short of significance. However, in situations such as the present one, where statistical power is low because of small group sizes, it is advisable to evaluate group differences by using an estimate of the magnitude of the mean difference, the "effect size", as well as by its statistical reliability (Cordray, personal communication). Given current guidelines (e.g., Cohen, 1977), the effect size for the LD vs N9 comparison can be said to be large (- .73), while that for the LD vs N4 comparison is small (.16) (see Table 24).

A parallel set of analyses of performance during the second test construction phase (TC2) yielded results comparable to those obtained for the first phase. The ANOVA indicated significant overall differences among groups, $F(2,33) = 5.51$, $p < .05$, but planned comparisons fell short of significance. The effect sizes again indicate a large difference between the LD and N9 groups (- .83) and a smaller difference between the LD and N4 groups (.46).

An inspection of means within and across groups (Table 22) suggests that there was a significant increase in the use of unconfounded tests from the first to the second test construction phase and that the improvement varied across groups. Tests of these trends were conducted using an analysis of covariance on final performance (TC2) with initial performance (TC1) as the covariate (see Table 21). Results revealed significant improvement across test phases,

$F(1,32) = 13.02$, $p < .05$, but no significant differential improvement among groups, $F(2,32) = 2.18$, $p > .05$.

Analysis of the more structured test choice and test evaluation phases of the rods task produced information concerning group differences which was generally consistent with that described above (see Table 23). There were significant group differences for the test evaluation questions, $F(2,33) = 6.45$, $p < .05$. The ninth grade controls performed significantly better than the LD and the fourth grade normal subjects, who did not differ. Mean differences on the test choice questions did not reach significance, $F(2,33) = 1.77$, $p > .05$, but the pattern of performance was similar to that for the other measures.

Comparisons With Same Age and Younger Controls: LV vs. N9 vs. N4. - The ANOVA on performance during the first test construction phase (TC1) indicated no significant differences among the groups in initial tendency to conduct unconfounded tests, F (2,33) = 2.29, $p > .10$ (see Table 25). However, a parallel analysis of TC2 data indicated significant group differences, F (2,33) = 5.82, $p < .05$. Multiple comparisons revealed that the fourth graders conducted a significantly smaller proportion of unconfounded tests ($\bar{X} = .25$) than either the normal-achieving (.54) or the low verbal ninth graders (.45), who did not differ (see Table 26).

A test of differential improvement in unconfounded testing among the three groups was conducted using an ANCOVA on the TC2 scores with the TC1 scores as the covariate. The analysis revealed significant improvement across groups, F (2,32) = 19.54, $p < .05$), and a marginally significant differential improvement among the groups, F (2,23) = 3.12, $p = .058$ (Table 25). Multiple comparisons indicate that this tendency is due primarily to a significant difference between the ninth and fourth grade control groups, with the older subjects showing more improvement than the younger subjects. The improvement of the low verbal subjects fell between that of the two control groups.

Analyses of the more structured test evaluation phase revealed a different pattern of performance than that seen during test construction phases. There were significant differences among the groups, F (2,33) = 9.85, $p < .05$; however, unlike the test construction phases, the low verbal group scored more like the fourth grade controls than like their ninth grade peers (see Table 27). Multiple comparisons tests indicated that the ninth grade controls (4.12) answered correctly significantly more questions than either the fourth grade controls (2.10) or the low verbal subjects (1.90), who did not differ. Although this same pattern of performance was evident in the means from the test choice phase, the differences were not significant, F (2,33) < 1.

Comparisons With Same Age and Younger Controls: LP vs. N9 vs. N4. - The ANOVA on performance during the first test construction phase (TC1) (Table 29) indicated no significant differences among the three groups, F (2,30) = 2.46, $p > .100$. However, an examination of effect sizes (Table 32) indicates that the difference between the two ninth grade groups (N9 and LP) is small (-.20), while the difference between the fourth grade group (N4) and each of the older groups is large (.92 and .89 for LP and N9 respectively).

This pattern of relative performance is again evident in the analysis of the data from the second test construction phase (TC2) (Table 29). There was a

significant difference among groups, F (2,30) = 5.90, $p < .05$. Multiple comparisons (Tukey's HSD) indicated that the fourth grade group (.25) constructed a significantly smaller proportion of unconfounded tests than the ninth grade controls (.54) or the low performance group (.49), who did not differ significantly.

An analysis of differential improvement in unconfounded testing from TC1 to TC2 was conducted using an ANCOVA on TC2 scores with TC1 scores covaried (Table 29). The ANCOVA revealed significant overall improvement, F (1,29) = 16.34, $p < .05$, and marginally significant differential improvement across groups, F (2,29) = 2.91, $p = .07$. Multiple comparisons revealed no significant pairwise differences between groups; however, an inspection of effect sizes (Table 32) indicates a pattern of group differences similar to that seen in the previous analyses. The differences between the fourth grade controls and the two ninth grade groups are large, while the difference between the two ninth grade groups is small.

The same pattern of group differences was again evident in the scores from the more structured test choice and test evaluation phases (Table 31). There were significant differences among groups for the test evaluation questions, F (2,30) = 6.40, $p < .05$. Multiple comparisons revealed that the fourth grade group answered significantly fewer questions correctly (2.10) than the low performance (4.00) or ninth grade controls (4.12), who did not differ.

Group means for the test choice questions were not significantly different, F (2,30) < 1, $p > .10$, but once again, effect sizes indicate "moderate" differences between the fourth grade control and the two ninth grade groups, and "small" differences between the two older groups.

Comparisons of Equivalent and Lower IQ Groups: ND vs. High vs. Low. - The ANOVA on performance during the first test construction phase indicated significant differences among the groups in initial tendency to conduct unconfounded tests, F (2,23) = 4.96, $p < .05$ (see Table 33). As can be seen in Table 34, this result is due to the fact that the High IQ group conducted a significantly higher proportion of unconfounded tests (.47) than the Low IQ (.22) and ND groups (.22), who did not differ. The analysis of performance during the second test construction phase produced similar results (see Tables 33 and 34). There was a significant difference among the groups, F (2,23) = 6.72, $p < .05$, and the High IQ group conducted a higher proportion of unconfounded tests (.67) than either of the other two groups, who did not differ (Low IQ: .36; ND: .39).

The test of differential improvement in unconfounded testing from TC1 to TC2 revealed significant overall improvement, F (1,22) = 15.40, $p < .05$, but no significant differential improvement across groups, F (2,22) = 2.31, $p > .10$. Examination of effect sizes, however, suggests that the same pattern of group differences is present in these data (Table 36). The difference between the ND and Low IQ groups is very small (.14), while the difference between each of these groups and the High IQ group is very large (-1.88 and 3.85, respectively). Thus, the High IQ group begins with the highest proportion of unconfounded tests and also shows the greatest amount of improvement.

Analysis of the more structured test choice and test evaluation questions revealed significant group differences, F (2,23) = 14.21, $p < .05$ and

F (2,23) = 4.16, $p < .05$ respectively (see Table 35). However, the pattern of differences among the groups varies. For the test choice questions, the relative performance is similar to that for the measures discussed above; however, for the test evaluation questions, the only significant difference is that between the High IQ (4.10) and the ND (2.40) groups. The performance of the Low IQ (3.67) group is similar to that of the High IQ group.

Comparisons of Equivalent and Lower IQ Groups: LV vs. High vs. Low. - The ANOVA on performance during the first test construction phase (TC1) indicated no significant differences among the groups in initial tendency to conduct unconfounded tests, F (2,23) = 2.28, $p > .10$ (see Table 37). However, a parallel analysis of TC2 data indicated significant group differences, F (2,23) = 6.18, $p < .05$. Multiple comparisons revealed that the High IQ ninth graders conducted a significantly higher proportion of unconfounded tests (.67) than either the LV (.45) or Low IQ (.32) groups, who did not differ (see Table 38).

The test of differential improvement in unconfounded testing from TC1 to TC2 revealed significant overall improvement, F (1,22) = 23.97, $p < .05$, and marginally significant differential improvement across groups, F(2,22) = 3.32, p = .055 (see Table 37). Post hoc tests indicated that the High IQ group showed significantly more improvement than the Low IQ group. The pattern of group means (Table 38) and the corresponding effect sizes (Table 40) indicate that the differences between the High IQ group and the other two groups are large, while that between the LV and Low IQ groups is small. Thus, the High IQ group begins with the highest proportion of unconfounded tests and also tends to show the greatest amount of improvement, as was the case with the comparisons to the ND group.

Analysis of the more structured test choice and test evaluation questions revealed significant group differences, F (2,23) = 9.36, $p < .05$, and F (2,23) = 8.76, $p < .05$, respectively. However, the pattern of differences among the groups varies (see Table 39). For the test choice questions, the relative performance is similar to that for the measures discussed above; however, for the test evaluation questions, the only significant difference is that between the High IQ (4.40) and the LV (1.90) groups. The performance of the Low IQ group (3.67) is similar to that of the High IQ group.

Comparisons of Equivalent and Lower IQ Groups: LP vs. High vs. Low. - The ANOVA on performance during the first test construction phase (TC1) indicated no significant differences among the groups in initial tendency to conduct unconfounded tests, F (2,20) = 2.61, p = .10 (see Table 41). However, a parallel analysis of TC2 data indicated significant group differences, F (2,20 = 5.54, $p < .05$. Multiple comparison tests revealed that the High IQ group conducted a significantly higher proportion of unconfounded tests (.67) than did the Low IQ group (.32). The performance of the LP group (.49) fell halfway between the other two groups and was not significantly different from either one (see Table 42).

The test of differential improvement in unconfounded testing from TC1 to TC2 revealed significant overall improvement, F (1,19) = 20.37, $p < .05$, but no significant differential improvement across groups, F (2,19) = 2.52, $p > .10$ (see Table 41). Examination of group means (Table 42) and effect sizes (Table 44), however, suggests that the same pattern of group differences is

present in these data. The difference between the High and Low IQ groups is large, while the differences between the LP group and the two control groups are moderate in size and equal. As in the comparisons of the other two LD subgroups to the two IQ control groups, the High IQ group begins with the highest proportion of unconfounded tests and shows the greatest improvement. The Low IQ group, in contrast, has the lowest scores and the least improvement, while the LD group falls in between.

Analysis of the test choice questions revealed significant group differences, F (2,20) = 11.42, $p < .05$. Post hoc tests indicate that the Low IQ group answered significantly fewer questions correctly (1.17) than did the LP (3.20) or High IQ (4.30) groups, who did not differ (see Table 43). The same relative group performance was evident in the test evaluation question data, but there were no significant group differences, F (2,20) < 1, $p > .10$.

Summary

The results of the analysis of group differences in preference for unconfounded tests are complex, but there are some general patterns which can be abstracted. In describing these patterns, the results of both the statistical tests and of the inspection of effect sizes will be used.

Although there were no statistically significant differences among the three LD subgroups in their tendencies to construct unconfounded tests, there was a consistent trend favoring the two high discrepancy groups. The performance of the LV and LP groups was roughly comparable and better than that of the ND group. On the more structured test choice and test evaluation questions, the LP group maintained a strong preference for unconfounded tests, but the LV group showed relatively lower performance and actually performed below the ND group on the test evaluation questions. This drop in relative performance may have been a result of greater verbal demands. For both the test choice and test evaluation questions, a subject's success is in part dependent on adequate comprehension of a complexly-worded question and on the ability to formulate an acceptable response. While many subjects in the ND group have some language difficulties, as evidenced by the teacher questionnaires, they may also have more higher-level cognitive difficulties.

The difficulties of the ND subjects are made more apparent by comparisons of their performance with that of normals varying in age and in fullscale IQ. In a comparison with same-age and younger controls, the ND group performed significantly below the level of the same-age control group and at a level equivalent to the younger (fourth grade) group. This relative performance was evident on all measures of preference for unconfounded testing. In the comparison to same-age controls varying in IQ level, the ND group performed no differently than the lower of the two IQ groups, despite the fact that this group had an average fullscale IQ slightly below that of the ND group. The one exception to this pattern was on the test evaluation questions, where the ND group performed significantly worse that the low-IQ control group.

The relative performance of the LP group was equally clear across analyses, but this group showed a pattern of performance which was the opposite of that seen in the ND group. In comparisons with same-age and younger controls, the LP subjects scored similarly to the same-age group and significantly above the level of the younger controls on all measures of preference for unconfounded

testing. They also showed a tendency to improve within the session comparable to that of the same-age controls. In comparisons to the two IQ-level control groups, the performance of the LP subjects fell midway between that of the high and low IQ groups.

The pattern of relative performance of the LV group is more complex than that of the other two LD groups. In the comparison with same-age and younger controls, the LV subjects performed at the same level as the same-age subjects on measures of test construction, but they performed more like younger controls on the test choice and test evaluation questions. The LV subjects performed at a level between the high and low IQ controls on the test construction and test choice measures. On the test evaluation questions, however, they performed below both IQ groups.

In general, the two high discrepancy groups showed a consistently greater preference for unconfounded tests than did the ND group. They performed at a level equivalent to same-age controls on test construction measures, while the ND group performed more like fourth grade controls. In comparisons to high and low IQ groups of the same age, the LV and LP groups performed between the two comparison groups, while the ND group performed more like the low IQ group. On the more structured and more verbal test choice and test evaluation measures, the LP group maintained its relatively high performance, but the LV group performed more like the ND group than the LP group. Thus, the three LD groups showed a differing pattern of performance, with two of the three groups looking much like normal controls on several measures. However, the pattern of preference for unconfounded tests varied somewhat as a function of the task demands.

Comparisons of Performance on the Rods and Conductivity Tasks

The conductivity task was included during the Assessment Phase for two reasons. First, in order to have some assurance that the performance observed on the rods task was a reflection of differential mastery of the isolation of variables strategy and not a trivial function of the particular features of the rods task, it was necessary to assess performance on a second task requiring the same strategy. In this sense, the conductivity task was used to estimate the generality of the findings obtained with the rods task. The second purpose for including the conductivity task relates to the Intervention Phase. Since a task amenable to multiple presentations was required for the intervention sessions, the rods task was inappropriate. If a different task was to be used, then some data on its comparability to the task used during the Assessment Phase would be necessary.

Two approaches were used to assess comparability of the rods and conductivity tasks. First, level of performance on each task was assessed. Second, the relative sensitivity of the tasks to differential performance was examined. These analyses will be summarized in the following sections. In reviewing the data to be presented it should be remembered that the order of presentation of the two tasks was not counter-balanced. Since the rods task was of fundamental importance for the Assessment Phase, it was always given first in order to avoid possible contaminating effects of the order of presentation. This decision necessitated a confounding of presentation order and task, however, which clouds any conclusions to be drawn.

Level of Performance

Table 45 contains data relevant to the issues of relative level of performance on the two tasks. Two aspects of the data contained in the table are of interest. First, it should be noted that the proportion of unconfounded tests on the conductivity task is higher than that for the initial test construction phase of the rods task but very comparable to that for test construction phase 2. Whether this fact reflects maintenance of the improvement evidenced during the rods task on a similar task of comparable difficulty, or initially higher performance on a task which is somewhat easier than the rods is impossible to determine.

Sensitivity to Differential Performance

Another index of the comparability of the rods and conductivity tasks comes from a comparison of the results of a set of ANOVAs of differential group performance on the conductivity task with those results obtained from the comparable rods task ANOVAs. These results are summarized in Tables 46 and 47. Results of statistical tests (Table 46) suggest that the conductivity task yields group differences similar to those found with the data from the initial test construction phase of the rods task. There are no significant differences for any of the group comparisons except the comparison of the ND group with same age and younger controls. The patterns of group means for this one comparison are identical for those two sets of data (see Table 45).

While the results of the analyses of the data from the second test construction of the rods task revealed more significant group differences than were evident in the conductivity data, an inspection of the group means (Table 45) and of the effect sizes (Table 47) suggest that the pattern of data from TC2 is actually more comparable to that found with the conductivity task than is the pattern from TC1. These results suggest once again that the behavior ultimately elicited during the second test construction phase is maintained on the conductivity task.

Another approach to assessing the comparability of the two tasks as measures of use of the isolation of variables strategy involves the creation of strategy-status groups on the basis of performance on the rods task. The relative performance of these groups on the conductivity task can then be evaluated. For this analysis, the criteria for the strategy-status groups were as follows. Subjects were first assigned scores for performance during TC1 and TC2 based on the number of the five relevant variables (length, diameter, material, weight, and base) which wre tested in a predominantly unconfounded manner. Then, subjects who received scores of 3 or more on this new measure during TC1 were said to have "spontaneous" access to the IV strategy. Subjects who met this criterion during TC2 but not TC1 were said to have "elicitable" access to the strategy. Finally, the IV strategy was said to be "absent" in subjects who failed to meet the criterion during either TC administration. Table 48 contains the results of this analysis of strategy-status groups. As can be seen, there is a significant difference in conductivity performance among the rods strategy-status groups, $F(2,50) = 5.44$, $p < .05$. This result reflects the fact that the strategy-absent group performs a significantly smaller proportion of unconfounded tests (.33) than either the elicitable (.50) or the spontaneous (.55) groups, which do not differ. Once again, these data suggest that

ultimate levels of performance on the rods task are maintained on the conductivity task one week later.

Another sensitive index of task comparability is the Pearson correlation. Correlations of the unconfounded testing scores from the rods task with that from the conductivity task are displayed in Table 49. The table contains both zero-order correlations and partial correlations controlling for fullscale IQ. As can be seen all correlations are significant, even with IQ partialled out. It should be noted that, consistent with the above analysis, the relationship is stronger for TC2 than for TC1. All correlations, are, however, only moderate in size.

Summary

The overall impression gained from the analyses summarized above is that there is a fair degree of comparability between the rods and conductivity tasks, both in terms of level of performance and in terms of sensitivity to individual and group differences. Also, it should be noted that these conclusions are more true for the second than for the first test construction phase of the rods task. This fact suggests that performance on the conductivity task represents the maintenance across task differences of the final level of performance from the rods task.

Comparison of Rods and Conductivity Tasks to Woodcock-Johnson

In order to obtain a preliminary estimate of how the measures of reasoning and problem solving obtained from the rods and conductivity tasks relate to currently available standardized measures of "thinking skills," the Concept Formation subtest of the Woodcock-Johnson Psychoeducational Battery (Woodcock & Johnson, 1977) was administered to each subject at the end of the final session. Standard admininstration and scoring procedures were used.

Table 50 contains a summary of the correlations between the Concept Formation (CF) standard scores and the measures of unconfounded testing from the rods and conductivity tasks for the ninth grade subjects. Both zero-order correlations and partial correlations, controlling for WISC-R fullscale IQ, are included. As can be seen, there are significant correlations among measures, but they are modest, representing between 8 and 25 percent shared variance. The strongest relations between the experimental measures and the CF task are found for the more structured test choice and test evaluation questions on the rods task, and for the measure of unconfounded testing on the conductivity task. Correlations between initial test construction performance on the rods task and CF scores are not significant. All correlations are reduced somewhat when the common variance attributable to fullscale IQ is partialled out. It is interesting to note that these reductions are largest for measures taken from early portions of the rods task (test construction 1 and test choice).

As a whole, these correlations suggest that there is a significant, but modest, relationship between unconfounded testing measures and a measure of nonverbal conceptual skills. This relationship becomes stronger as variance due to initial rods task ambiguity is lessened. However, there is still a large proportion of variance which is unique to the experimental tasks.

Group Differences in General Task Approach

Overview of Analyses and Technical Considerations

The analyses to be reported in this section were intended to provide information about group differences in task approach (See Objectives 4 and 5). The goal of these analyses is to provide a detailed description of subjects' behavior during both phases of the administration of the rods task. The behaviors analyzed were selected from a larger number in a general coding system (see Appendix B) designed to index four broad dimensions of performance in a problem solving task. These four dimensions are goal setting, data gathering, reasoning, and need for examiner guidance. The measures implemented here were chosen because of their likely potential for differentiating subjects' task approach. These measures were analyzed using a multivariate discriminant function procedure in order to provide a composite behavioral description of group differences in task performance.

Six indices of general task approach were analyzed (see Dependent Variables). Three of these six variables, articulation, generality, and use of explanatory principles, assessed the degree of sophistication of subjects' conclusions. Two of the variables, intentionality and explicitness, indicated the nature of their predictions. The remaining variable, noncompliance with examiner guidance, measured the subjects' ability to produce tests of a specific variable upon request.

In order to derive a rich, multifaceted description of group differences in task approach, the six variables were analyzed using discriminant function analyses. Current guidelines for the use of discriminant analysis suggest a subject to variable ratio on the order of 10 to 1. However, given the exploratory and descriptive nature of the present study and the expense involved in collecting additional data, it was decided to use a less stringent criterion (a ratio of 5 to 1). A multivariate descriptive procedure, such as discriminant analysis, was selected in order to control for the intercorrelations among these variables.

The following comparisons of task approach among the six groups were made: the three LD subgroups were contrasted, and each LD subgroup was compared, in turn, with the older and younger normal-achieving groups. Contrasts between the LD subgroups and low and high IQ ninth grade controls were not analyzed due to the small sample size. Given the absence of any normative data on age differences in the constellation of variables included in these analyses, specific planned comparisons such as those used in the preceding analysis of unconfounded testing were not possible. A separate discriminant analysis was run for each test construction phase.

The discriminant function analysis provides two kinds of information. First, it provides an indication of which variables best differentiate the subject groups and how well these variables collectively discriminate among the groups. Second, it computes a discriminant function score for each subject which can be used to predict that subject's most likely group assignment given his or her scores on the original set of variables. The procedure's ability to estimate accurately each subject's group membership is another index of the discriminating power of the function. Two summary indices of classification success will be reported: the proportion of cases correctly classified when

the whole sample was used to generate the function and the proportion correctly classified when each case was eliminated in turn using a "jackknife" procedure. The jackknife procedure enables one to estimate the classification success rate expected over all possible future samples of a given size (Huberty, 1984). In studies where small sample sizes can affect the stability of the estimate of success rate, the jackknife procedure yields a lower bound on the replicability of the discriminant function.

A detailed description of each of the four sets of discriminant function analyses is provided below. Following the detailed presentation of the results is a summary of the general findings.

Group Differences in Task Approach Among Three LD Subgroups

Test Construction Phase 1. - Inspection of the results of the discriminant function analysis for the first test construction phase indicates that neither the first nor the second function was capable of discriminating between the three LD subgroups at a statistically significant level. The canonical correlation between the first function and a dummy-grouping variable is .463, which indicates that 21% of the variance among the subject groups is explained by this one discriminant function. The first function resulted in a Wilk's Lambda of .662 (χ^2 = 8.87, df = 12, p = .714). The second function generated by the analysis produced a canonical correlation of .397, which means that 16% of the variance among the groups is explained by this second function. (Wilk's Lambda = .842, χ^2 = 3.69, df = 5, p = .595.)

The group centroids indicate that the first function discriminates the LP and LV groups with the ND group scoring somewhere in the middle (see Table 52). The standardized coefficients show that the two variables with the highest coefficients (in absolute value) are intentionality and explicitness of predictions with the LV group making less explicit predictions about their tests and showing less of a tendency to express an intent in their predictions (see Table 51). These two variables contribute the greatest explanatory power to this first function.

The second function discriminates the ND group from the other two learning disabled groups. The standardized coefficients show that the one variable that contributes the most to the second function is generality of conclusions with the ND group using more general conclusions than the other two learning disabled groups. The results of the classification procedure show that these two functions correctly classified 52% of the cases when the whole sample was included in the calculation of the functions, but only 30% of the cases when each case was eliminated in turn using a "jackknife" classification procedure. The relatively large drop in classification success in the present case seems to indicate that within-group variability is playing an important role in the functions' classification accuracy (see Table 53).

Test Construction Phase 2. - An identical set of procedures was performed for test construction phase 2 using the same six indices of performance. As in test construction phase 1, this second discriminant analysis produced no statistically significant functions. Thirty-four percent of the variance among the groups is explained by the first function (canonical correlation = .584, Wilk's Lambda = .622, χ^2 = 10.22, df = 12, p = .597)

and 6% of the variance is due to the second function (canonical correlation = .236, Wilk's Lambda = .944, χ^2 = 1.24, df = 5, p = .941).

The first function again discriminates the LV group from the LP group, based on their centroids, with the ND group's centroid falling in between (see Table 55). The variable with the largest standardized coefficient on function 1 is articulation of conclusions. This indicates that the LP group supplies more highly articulated conclusions in test construction phase 2 than does the LV group (see Table 54). In addition, the LP group is also more likely than the LV group to use explanatory principles in their conclusions and to indicate an intention to test a particular hypothesis in their predictions.

The second function differentiates the performance of the ND group from the other two. Like the LV group, the ND subjects express fewer intentions to test hypotheses in their predictions than do the LP subjects. In contrast, like the LP group, the ND subjects use more explicit predictions than do the LV subjects.

The results of the classification procedure for test construction phase 2 indicate that these two functions correctly classified 63% of the cases when the whole sample was included, but only 33% of the cases when each case was eliminated in turn using the jackknife procedure (see Table 53).

Comparisons with Same-age and Younger Controls: LV vs. N9 vs. N4

Each of the three LD subgroups was compared, in turn, with the groups composed of same-age and younger controls using discriminant function analyses. Again, a separate analysis was run for each test construction phase.

Test Construction Phase 1. - The results of the discriminant function analysis for test construction phase 1 indicate that neither of the two functions was statistically significant. The canonical correlation between the first function and a dummy-grouping variable is .558, which indicates that 31% of the variance among the three groups is explained by this first function. (Wilk's Lambda = .572, χ^2 = 15.94, df = 12, p = .194.) The second function explained 17% of the variance. (Canonical correlation = .412, Wilk's Lambda = .830, χ^2 = 5.29, df = 5, p = .381.)

The group centroids indicate that the first function discriminates the two normal control groups. The LV group centroid falls midway between the centroids of the two control groups (see Table 56). The three variables with the highest standardized coefficients are: use of explanatory principles, generality of conclusions and noncompliance with prompts. This first function indicates that the normal ninth graders use more explanatory principles, make more general conclusions, and are more likely to comply with the examiner's requests for particular tests than are the normal fourth graders (see Table 51). The LV ninth graders are more like their normal age-mates in the generality of their conclusions but are less likely than the N9 group to use explanatory principles. In addition both the LV and N4 groups appear to have more trouble complying with the examiner's requests.

The second function discriminates the LV group from both of the normal control groups. One variable, intentionality of predictions, contributes the primary discriminating power of this function. It shows that the LV group is less

likely than either normal group to express an intention to test a particular hypothesis in their predictions.

The classification procedure for test construction phase 1 indicates that these two functions correctly classified 68% of the cases when the whole sample was included, but only 35% of the cases using the jackknife procedure (see Table 57).

Test Construction Phase 2. - The discriminant function analysis for test construction phase 2 produced two functions, the first of which approached statistical significance at the p =.10 level. The first function accounted for 41% of the variance among the three groups (canonical correlation = .639, Wilk's Lambda = .532, χ^2 = 17.99, df = 12, p = .116). The second function explained 10% of the variance (canonical correlation = .318, Wilk's Lambda = .899, χ^2 = 3.04, df = 5, p = .694).

The first function discriminates the LV group from their normal age-mates (see Table 58). This function is defined by three variables, generality of conclusions, noncompliance with prompts and articulation of conclusions. It indicates that the LV ninth graders make more general but less highly articulated conclusions than do normal ninth graders. In addition, the LV group is less likely than the N9 group to comply with the examiner's request for specific tests (see Table 54).

The second function differentiates the LV and N4 groups. This function is defined by a single important variable, generality of conclusions. It shows that the younger controls make more specific conclusions about their tests than do the LV ninth graders (see Table 54).

These two functions correctly classified 65% of the cases using the entire sample and 41% of the cases using the jackknife procedure (see Table 57).

Comparisons with Same-age and Younger Controls: LP vs. N9 vs. N4

Test Construction Phase 1. - The discriminant function analysis produced two functions, neither of which was statistically significant. The first function explained 38% of the variance between the groups (canonical correlation = .621, Wilk's Lambda = .573, χ^2 = 14.21, df = 12, p = .288). The second function accounted for 7% of the variance (canonical correlation = .259, Wilk's Lambda = .933, χ^2 = 1.77, df = 5, p = .880).

The first function discriminates the two normal control groups (see Table 59). Note that the LP subjects score closer to their age-mates than to the fourth graders on this function. The standardized coefficients indicate that two of the variables, use of explanatory principles and generality of conclusions, contribute the greatest explanatory power to this function. As we have seen in previous comparisons (e.g., LV vs. N9 vs. N4), the normal ninth graders tend to employ more explanatory principles and to make more general conclusions than do the fourth graders. In this comparison, the LP ninth graders perform, on both of these measures, more like their normal age-mates (see Table 51).

The second function discriminates the LP group from their normal age-mates. The two defining variables for this function are explicitness of predictions

and noncompliance with prompts. This analysis indicates that LP ninth graders make more explicit predictions about their tests but are less likely to comply with the examiner's request for specific tests than are the normal ninth graders (see Table 51).

These two functions correctly classified 61% of the subjects using the entire sample and 42% of the subjects using the jackknife procedure. Inspection of the misclassifications using the standard or whole sample procedure (see Table 60) reveals that 71% of the LP sample was misclassified as belonging to the ninth grade normal group based on their function scores. This finding suggests that the verbal reasoning of the LP group during test construction phase 1 is quite similar to that of their normal age-mates.

Test Construction Phase 2. - The discriminant function analysis for the second test construction phase produced two functions, neither of which was statistically significant. The first function accounted for 26% of the between group variance (canonical correlation = .513, Wilk's Lambda = .657, χ^2 = 10.69, df = 12, p = .555). The second function explained an additional 11% of the variance (canonical correlation = .329, Wilk's Lambda = .892, χ^2 = 2.92, df = 5, p = .713).

The group centroids for the first function indicate that it differentiates the fourth grade group from both ninth grade groups (see Table 61). As in test construction phase 1, the LP group performs more like their normal age-mates than like the younger controls. The two defining variables for this function are noncompliance with prompts and articulation of conclusions. Inspection of the group means for these variables in Table 54 shows that both the LP and N9 groups are making more highly articulated conclusions and are more adept at producing the tests requested by the examiner than are the fourth graders.

The second function discriminates the two ninth grade groups. The two important variables in this function are explicitness and intentionality of predictions. This function reveals that the LP ninth graders are producing less explicit predictions than are their normal age-mates. However, the LP group is more likely to express an intention to test an hypothesis than is the N9 group.

The results of the classification procedure for test construction phase 2 show that 58% of the subjects were correctly classified using the whole sample, but only 39% were correctly classified when the jackknife procedure was employed. As in test construction phase 1, a substantial number of misclassifications using the standard procedure occurred in the LP group (see Table 60). Fifty-seven percent of the LP group were assigned to the N9 group based on their function scores. However, a large proportion of the normal fourth graders (40%) would have also been assigned to the N9 group.

Comparisons with Same-age and Younger Controls: ND vs. N9 vs. N4

Test Construction Phase 1. - The first of the two functions produced by the discriminant analysis procedure approached statistical significance at the p = .10 level (Wilk's Lambda = .546, χ^2 = 17.27, df = 12, p = .140). This first function accounts for 41% of the variance between the groups (canonical correlation = .642). The second function is not a statistically significant

discriminator and explains 7% additional variance (canonical correlation = .268, Wilk's Lambda = .928, χ^2 = 2.13, df = 5, p = .830).

The first function discriminates the fourth grade group from the two ninth grade groups whose average function scores are quite similar (see Table 62). The two defining variables for this function are generality of conclusions and use of explanatory principles. Inspection of the group means on these variables (Table 51) reveals that the two ninth grade groups make more general conclusions and use more explanatory principles than do the fourth graders.

The second function differentiates the ND and N9 groups. The two important variables for this function are generality of conclusions and noncompliance with prompts. This function reveals that the ND ninth graders make more general conclusions than their normal age-mates, but are less likely to comply with requests from the examiner for specific tests.

The classification procedure correctly classified 56% of the subjects when the whole sample was used and 38% when the jackknife procedure was employed. Inspection of the misclassification table (Table 63) reveals that both procedures misassigned a majority of the ND group to the N9 group (60% using the standard procedure, 80% using the jackknife procedure). Also, a sizeable minority of the N9 group were misclassified as ND (33% using the standard procedure, 36% using the jackknife procedure). None of the fourth graders was misclassified as ND. This suggests that there is a substantial overlap in performance on the variables examined here between the subjects in the two ninth grade groups.

Test Construction Phase 2. - The discriminant function analysis produced two functions, neither of which was statistically significant. The first function accounted for 25% of the variance between the groups (canonical correlation = .498, Wilk's Lambda = .651, χ^2 = 12.22, df = 12, p = .428). The second function explained an additional 13% of the variance (canonical correlation = .365, χ^2 = 4.08, df = 5, p = .538).

The first function discriminates the normal ninth graders from the ND and fourth grade groups whose average functions scores are similar (see Table 64). The defining variables for this function are articulation of conclusions and noncompliance with prompts. Table 54 shows that the normal ninth graders are supplying more highly articulated conclusions and are more successful at responding appropriately to the examiner's requests for tests than are either the ND or fourth grade groups.

The second function helps differentiate the ND ninth graders from the fourth graders. The most important variable for this function is generality of conclusions. It indicates that ND ninth graders tend to make more general conclusions about their tests than do fourth graders.

Fifty-nine percent of the sample was correctly classified by the function when the entire sample was used and 44% was correctly classified using the jackknife procedure.

Summary

What were the major differences in task approach among the three LD subgroups and between each LD subgroup and the normal controls? The discriminant analyses revealed that the differences that exist among the groups are not dramatic nor are they significant in a statistical sense. However, the analyses did reveal some consistent trends in the data.

With respect to the three LD subgroups, the LP was the strongest in verbal reasoning. Initially, the LP group showed this strength in their predictions about their rod selections. Their predictions were both more explicit and revealed more of an intention to test either inductive or deductive hypotheses than were those of the LV group. During test construction phase 2, the LP group maintained their superiority in reasoning over the LV group. However, this time it was the conclusions of the LP group that were better. The LP group were making more highly articulated conclusions and were more likely to use explanatory principles than the LV group. On the average, the verbal reasoning of the ND group fell in between that of the other two LD subgroups during both phases of the task. This finding may not be surprising, if one takes into account the relatively higher average verbal IQ of the LP group (= 109.7) in comparison with that of the ND (= 97.5) and LV (= 88.8) groups. It appears then that, at least with respect to the aspects of verbal reasoning assessed by these measures, the relatively poor visual-spatial abilities of the LP subjects did not interfere with their ability to hypothesize and express explicit predictions or with their skill at articulating conclusions or providing occasional explanatory principles.

When the LV subgroup was compared with same-age and younger controls, the following differences in task approach were observed. During the initial administration of the task, the two normal control groups were the most dissimilar. The normal ninth graders supplied more general conclusions and used more explanatory principles than did the normal fourth graders. The normal ninth graders also showed their greater understanding of the task principles by supplying more of the examiner requested tests. The LV ninth graders were most unlike the other two control groups in their inability to express an intention to test the effects of a particular variable. However, by test construction phase 2, the LV group was approaching the task quite differently from either normal control group. While the normal ninth graders were making quite specific but well articulated conclusions about their tests in TC2, the LV group's conclusions were less well articulated and overly general. What this suggests is that the conclusions of the LV subjects were lacking in specificity: They weren't exploring the effects of one variable on another (e.g., length in steel rods) nor were they making explicit the continuous functional relationship between an attribute and the outcome. In addition, the LV group had more trouble understanding the examiner's requests for particular tests than did their normal peers. The LV group differed from the fourth graders in their continued use of general conclusions. The fourth graders tended to make specific conclusions during both test construction phases while the normal ninth graders began making general conclusions, then switched to specific conclusions during TC2.

Again, the verbal strengths of the LP group are apparent when they are compared with the two control groups. During both phases of the task, the LP group performed more like their normal age-mates than like the fourth graders.

For TC1, both ninth grade groups differed from the fourth graders in their use of explanatory principles and general conclusions. The major differences between the two ninth grade groups during the initial phase of the task were in the more explicit predictions of the LP subjects and in their greater difficulty complying with examiner-requested tests. During TC2, the LP and N9 groups showed their greater understanding of the task demands by supplying more highly articulated conclusions and by complying more often with examiner prompts than did the fourth graders. The two ninth grade groups differed in their predictions. The LP group gave less explicit predictions but were more likely to express an intention to test an hypothesis than did the N9 group.

When the ND group was compared with the two control groups, they performed more like their age-mates initially, but more like the fourth graders during TC2. During TC1, the two ninth grade groups drew more sophisticated conclusions about their tests than did the fourth graders. In particular, they were more likely to express general conclusions and to employ explanatory principles. The ND group differed from their normal age-mates in the generality of their conclusions and in the inability to comply with examiner prompted tests. Thus, although both ninth grade groups made general conclusions about their tests, the ND ninth graders were the most general. The ND group had more difficulty than the N9 group comprehending the requests of the examiner for particular tests.

By TC2, the ND group looked more like fourth graders. The major difference between the younger controls and ND ninth graders on the one hand and the normal ninth graders on the other were articulation of conclusions and noncompliance with examiner prompts. This difference indicates that the normal ninth graders could demonstrate their mastery of the task demands through both verbal reasoning and comprehension of complex task instructions. Unlike the fourth graders, the ND ninth graders continued to supply highly general conclusions for their tests. Both ND and LV subgroups showed this tendency to give overly general conclusions during TC2. This tendency may suggest an inability to learn from the feedback provided by their prior activities. In contrast, normal ninth graders tended to supply more specific conclusions during TC2: an indiction that they may be exploring the effects of variables on subsets of rods (e.g., diameter in plastic rods) because they have already established that each of the variables affects bending in general.

Development of Teacher/Clinician Rating Scale

One goal of the project (Objective 7) was the development of a modified assessment procedure for identifying reasoning problems in learning-disabled adolescents that could be used by LD clinicians in the field. The Teacher Rating Scale (TRS) which evolved in pursuit of that goal is designed for use with the rods task, in particular the portion of the task described elsewhere as test construction 1. In view of the investment in materials and training, it is likely that the full task will not be used often by the general LD practitioner. However, the TRS can serve as a model for analogous scales to be used for describing adolescent behavior in approaching other types of reasoning tasks. A manual for the use of the TRS is attached as Appendix C.

The original suggestions regarding categories of behavior to include in the scale were made by experienced LD clinicians who already had some experience

with the formal coding system and who had already been involved in coding the project videotapes. Early versions of the scale and the manual were used to acquaint graduate students with the ongoing work of the project. Eventually, after changes and refinements suggested by experience with other forms of the scale, the present version was established, which involves the coding of the following six categories of behavior:

1. Rod Selection: Was the subject's selection of rods for testing "quick," with no apparent plan (or with a stimulus-bound plan such as serial order in the array), or was it "thoughtful," as evidenced by a selective search for a particular rod or pair of rods for testing, presumably based on rod attributes. This category, like number 4 ("variables mentioned"), is designed primarily to pick up one unusual approach to the task. Most subjects make "thoughtful" searches for rods at the beginning of each test, at least after the first one or two tests, but an occasional subject will test whatever two rods come to hand, or will arbitrarily test the two rods furthest to the left in the array, then the next two, etc. This approach to the task, when it occurs, is potentially diagnostic for severe thinking-skill deficits.

2. Confounded/Unconfounded Judgment: This category corresponds to that used in the formal coding, except that only the codes Confounded, Unconfounded, and Confounded by Base were used. The consideration of base proved difficult to convey to the raters who tested the scale. It is possible that the elimination of the base variable would considerably simplify the administration and coding of the task without sacrificing a great deal of information.

3. Reasons: For each test, the task administrator asks "Why did you pick those rods?" and the rater codes the subject's response according to the degree and type of intention expressed. This category of behavioral data gives a good insight into the subject's understanding of the purpose his task.

4. Number of Variables Mentioned: There are five salient variables in the rods task: length, diameter, material, base, and size of weights used to test relative bending. Before a subject can effectively "control" these variables, they must be perceived and categorized as potential variables. This code captures the extent to which each subject is aware of these variables by noting whether or not the variables are mentioned by the subject.

5. Conclusions: After each trial, the subject responds to the question, "What did you learn about bending?" The subject's responses give an indication of whether his testing is directed toward the immediate stimulus ("This long plastic one's a lot more flexibler.") or whether he or she is attempting to abstract general principles from the tests. ("This shows plastic has got to bend the most.")

6. Explanation: Although no specific request is made, subjects sometimes offer an explanation, implied or explicit, for the differential bending they observe. ("This doesn't bend more because wood is sturdier.") This is another indication of the subject's purposiveness in executing the task. This was by far the most difficult category to

describe and to code, partly because of the absence of a specific probe designed to elicit this kind of response.

Once the six categories of information had been settled upon and described in a coding manual, raters were recruited to assess the extent to which the codes could be reliably implemented by relatively naive clinicians. Five persons not connected with the project, all first-year Ph.D. students in learning disabilities with substantial teaching experience in the field, attended a 90-minute training session in which the manual was reviewed and a demonstration tape was shown. They then independently coded test construction 1 on three of the project videotapes. In order to simulate real-time coding as closely as possible, raters were allowed to stop the tape during the coding (just as they might delay proceedings in the administration of the task in order to think through or catch up to a code), but they were not allowed to rewind the tape for a second viewing. The three tapes were selected from the ND subgroup at random. After the raters filled out rating sheets for the three subjects, their ratings were compared to the ratings on a "master" code sheet which had been prepared by project members who had access to repeated viewings/hearings of the videotapes and to typed transcripts of the tapes. Tables 65 through 69 summarize the amount of agreement between the coding of the raters and that of the master code sheet, broken down by subject, by category of information, and by rater.

The overall accuracy of the raters was 70%. There was slightly more variance in agreement among the three cases (71%, 76%, and 63%) than there was among the five raters (range: 67% to 72%). One difficulty with this assessment of the TRS is that it probably takes more experience with the scale than is provided by three cases to internalize the categories of information required by the scale. Several of the raters estimated that they would not really "feel comfortable" with the scale until they had additional time to study the manual and had coded a dozen or more cases. The project data on the TRS reliability do not speak to this question; no increase in reliability was seen across the coding of three cases. However, this may be due in part to the fact that the case coded third by most of the raters was more difficult than the first two cases.

Variation among the six categories of codable information followed more or less predictable lines, with agreement highest on the more concrete scales, confounded/unconfounded (96%), rod selection (84%), and variables mentioned (83%), and lowest on the three scales which required judgments about the subject's linguistic output, reasons (45%), conclusions (57%), and explanations (54%). Examination of the data for the three latter codes revealed some consistent patterns in the disagreements. These patterns lead to suggestions for future revisions in the coding descriptions and criteria which might produce higher inter-rater reliabilities in future use of the TRS.

The major cause of disagreements for the Reasons code was the tendency of raters to assign 2's ("Description of variables only") to responses which the master coder had given 4's ("Intention to compare the effect of variables on bending"). An examination of the transcripts pertinent to several such miscodings suggests that this discrepancy occurred on tests in which the subject's intention was less than explicit. For example, in one test the subject responded to the "Why Pick?" question with "What would the plastic do compared to the metal ring." More explicit instructions regarding the

distinction between references to variables and references to testing variables would reduce this source of disagreement.

Disagreements on the Conclusions code took primarily the form of overestimates of generality on the part of the raters, i.e., assigning codes of 2 ("Generalized") when the master coder had assigned a 1 ("Single-test conclusion"). Inspection of the transcripts suggests that this trend is based in large part on internally inconsistent or underspecified utterances such as, "Just seeing how much stronger the metal is from a piece of plastic or glass." Further specification that only explicit and internally consistent general conclusions should receive high codes would reduce this source of error.

Inspection of the data from the Explanations code did not reveal a clear pattern of disagreements. At some times, the raters coded utterances as 2's ("Rod attributes as causal agents") when the master coder had assigned 1's ("None offered"). At other times, the opposite pattern was evident. Thus, there is no clear indication of how to improve inter-rater agreement on this code. Although this code has proved useful in discriminating groups in the present project (see Group Differences in General Task Approach), it may not be possible to implement it reliably in a real-time coding situation.

As a whole, the results of our testing of the Teacher Rating Scale are fairly encouraging. While inter-rater agreement was low on three codes, in two cases there is reason to believe that improvements could easily be made in the manual which would lead to better reliability. Also, it should be remembered that only a small number of tapes was coded and that there was variability across raters and tapes. Estimates of inter-rater reliability using a revised TRS and a larger data base would provide a better test of the ultimate utility of the scale.

INTERVENTION PHASE

Purpose and Overview

The intervention phase of the project was intended to provide a detailed description of the progress made by individuals exhibiting specific reasoning difficulties over a series of individually-designed instructional sessions. This phase had five guiding principles. First, it was to be exploratory and descriptive in focus. Second, the intervention was to be targeted to individuals, not groups of students. Third, a single-subject research design was to be used to assess the progress of individuals over time. Fourth, the intervention was to be modeled after a Piagetian clinical interview. Fifth, the aim was to document the learning process in the context of the IV task setting.

The rationale for this approach comes, in part, from a few multiple-session intervention studies of the development of the IV strategy in normal children (e.g., Kuhn and Phelps, 1982) (see Literature Review). The aim of these studies was to describe the changes in reasoning and data gathering strategies of individual subjects as they worked on IV tasks across a series of sessions. This research focused on observing behavioral change in a setting where the primary feedback about performance came from the task materials, not from the examiner. The examiner's activities were restricted to asking a standard set of questions of each subject in order to elicit

additional information about their problem-solving activities. Multiple-session intervention research with normal children has begun to provide detailed information about how individuals differ in their ability to organize and plan their problem-solving activities, to modify their behavior in response to feedback and to reason in a logical fashion. Thus, this research can supply a wealth of information about how individuals differ in their initial and subsequent task approach.

Like the multiple-session intervention research with normals, the intervention phase of the project focused on documenting the learning process of individuals within the IV task setting. However, in contrast to the research with normals, we allowed the examiner to use a wider range of probes in a few of the sessions. The intervention phase employed a single-subject research design and two IV tasks, the bending rods and a modified version of the chemical conductivity task. The bending rods task was used as a pretest to insure that the subjects of the intervention showed no evidence of the IV strategy spontaneously or after minimal prompting. The modified conductivity task was used over a series of sessions as both a baseline task and a focus for instructional activity. At the end of the intervention sessions, the bending rods task was readministered to assess overall improvement and generalization.

The examiner's behavior was monitored throughout the intervention phase. During the administrations of the bending rods task and the baseline sessions with the conductivity task, the examiner followed a script which was quite similar to that used during the assessment phase of the project. During the instructional sessions with the conductivity task, the examiner was instructed to restrict his or her remarks to a limited set of allowable prompts. At no time did the examiner attempt to teach the IV strategy in any explicit way. Instead, he or she tried to highlight inadequacies in the subjects' reasoning or data gathering strategies (e.g., "Could you tell <u>for sure</u> that electrode length mattered in these two bottles?") or to force the subjects to clarify and explain their ideas (e.g., "<u>Why</u> do you think bottle size doesn't affect brightness?").

The intervention phase of the project was intended to address objectives 8-11. However, our ability to meet these objectives was hampered by a number of factors. First, the body of relevant research is limited and has been restricted to normal-achieving children. Second, because the assessment phase of the project was delayed due to subject recruitment difficulties, we had an inadequate data base for intervention activities. Thus, we could rely upon neither previous research nor the results of our analyses of the assessment data for information concerning subgroups of reasoning and problem-solving difficulties among LD and normal adolescents. As a result, we decided to conduct a small-scale intervention study to explore the utility of a single-subject design and a Piagetian clinical interview instructional strategy.

The organization of this section is as follows. The first subsection contains the methodology used in this study. Included here are descriptions of the subjects studied, the tasks and procedures used, the intervention guidelines followed, and the coding procedures. The second subsection contains case studies of three of the LD adolescents who participated in this phase of the project. The case studies will include both qualitative and quantitative

information about their progress. In the final subsection, the findings are summarized, and preliminary suggestions for future research are discussed.

Methods

Subjects

The subjects who participated in the intervention phase of the project were recruited from one Chicago suburban school using a process similar to that employed in the assessment phase. Because of our primary interest in understanding the learning process in LD adolescents, we decided to restrict our small sample to ninth graders diagnosed as LD and receiving LD services. Thus, we did not collect data on normal controls during this phase of the project. The subjects used in this phase (all white males) were selected from a sample of ten subjects based on their WISC-R IQ profiles and on their performance during a double administration of the bending rods task. Subjects with a fullscale IQ at or below 85 were eliminated. (One subject with a fullscale IQ = 86 but with a verbal IQ = 75 was also eliminated.) The three subjects discussed in this section had a mean verbal IQ = 93.3 (range = 91-97), a mean performance IQ = 94 (range = 93-96) and a mean fullscale IQ = 93 (range 91-96). Subjects who demonstrated their ability to control 3, 4, or 5 out of 5 variables during either administration of the bending rods task were also eliminated.

Revised Conductivity Task

The materials employed were largely identical to those used in the assessment phase. (see Chemical Conductivity Task). In this revised version, each glass vial was labeled to indicate its contents. The labels contained one to three randomly selected letters such as H, J, B, HB, JH. Thus, some vials contained one chemical, some two, some three. In addition, the vials differed in two other ways: electrodes (either the length or the width between) and size (either the width of the vial or its shape).

In each session, the labels on the vials were different, and, occasionally, the vial size and electrode characteristics were different. However, the way the three chemicals interacted did not change. One chemical was quite effective in conducting electricity, the second chemical was a less effective conductor, and the third chemical did not conduct. When the strong conductor was paired with either the weak or nonconductor, its ability to conduct was impaired. The same was true for the weak conductor when paired with the nonconductor.

All subjects were presented with a set of 12 vials and the conductivity box. At the beginning of each session, the examiner demonstrated a test with two vials in which all five factors (3 chemicals, bottle size, and electrode length) were confounded. The examiner then asked the subject to explain why the light bulbs lit differently. Following the demonstration, the subject was allowed to construct his own tests as in test construction phase 1 in the bending rods task. After the subject indicated he was finished, the examiner returned to the initial demonstration pair and again asked for an explanation for the differential brightness.

The above procedure was followed during both types of chemical conductivity sessions: the baseline and the intervention sessions. In the baseline sessions, the examiner was scripted to three questions: why did you pick these? (after the vials were selected); what do you think will happen? (before the lights were lit); what did you learn? (after the lights were lit). However, in the intervention sessions, the examiner was allowed to use an additional set of prompts. (See below for a complete description of these additional prompts.)

Intervention Design and Guidelines

The intervention phase followed two general principles: a single-subject research design and a modified Piagetian clinical interview format. A single subject design was chosen because of its ability to document change in individuals who have been explosed to instructional activities. In addition, this design is useful when one is dealing with a heterogeneous population, such as LD adolescents, because of its ability to demonstrate individual differences in learning. A modified Piagetian clinical interview format was chosen because we were interested in investigating further the elicitability of the IV strategy in LD adolescents. Thus, we provided minimal examiner feedback to the subjects in order to observe how they responded to feedback from their own activities with the materials. More feedback from the examiner was provided during this phase of the project than during the assessment phase. However, the examiner's comments were focused on encouraging subjects to monitor their activities, to question their ideas, to justify their assertions, etc. At no time did the examiner attempt to teach the IV strategy. This indirect feedback also served two additional purposes: it provided us with more information about each subject's thinking and it indicated to the subject our interest in his ideas.

The single-subject research design employed had three components. The first component, the bending-rods task served as a pre-test, post-test and generalization measure. The second component was the baseline sessions with the conductivity task. These sessions preceded and followed the third component, the instructional sessions with the conductivity task. A typical design for an intervention with one subject could be characterized as follows: O_1 O_2 X X O_2 O_1 where O_1 = bending rods session, O_2 = baseline conductivity session, X = instructional conductivity session. Therefore, each subject was exposed to approximately six sessions. Although the majority of the sessions were used to gather baseline information about changes in reasoning and problem-solving, they, like the instructional sessions, exposed the subject to feedback from the task materials. In the instructional sessions, this feedback was augmented and highlighted by the examiner's prompts.

Before the intervention began, we identified six types of prompts that the examiner could use in the instructional session, in addition to the standard probes employed in the rods and baseline sessions. They were:

1. Non-standard probes - minor variations of the "standard" probes below;
2. Clarification - attempts by the examiner to get the subject to clarify a previous remark;
3. Justification - requests that the subject justify a previous prediction or conclusion;

4. Challenging understanding and goal structuring - challenges to the subject's current level of understanding and suggestions to the subject intended to help him keep the task goal in mind;

5. Reassurance - nonspecific encouragement;

6. Encouraging self-monitoring - requests that the subject reflect on his remarks or activities.

Prior to each instructional session, an instructional team composed of the examiner and one or two assistants decided which set of these prompts to emphasize. Decisions were based on the collective clinical judgments of the team who viewed together the videotapes of each subject's bending rods and baseline conductivity sessions. The decisions served as instructional guidelines for the examiner who tried to follow them in the subsequent session. However, in order to interact with the subject in a naturalistic manner, the examiner did not follow a strict script in the instructional sessions as he or she did in the baseline sessions.

Coding Procedures

Coding procedures for the intervention phase followed those developed for the assessment phase. These procedures yielded the same kinds of dependent variables (e.g., proportion of unconfounded tests, generality of conclusions) analyzed in the initial phase of the project (see Assessment Phase - Data Coding.). In addition, a code was devised to categorize the kinds of prompts used by the examiner during the instructional sessions. The Examiner Intervention Code is described below.

Examiner Intervention Code. - All clear, task-related requests for information, requests for specific tests, challenges to subject's understanding, reassurances, and suggestions by the examiner were coded. Answers to direct questions by the subject, informational remarks about the task materials or procedures, sentence fragments, unclear or ambiguous utterances were not coded. If the interpretation of an utterance depended upon the meaning of one or more preceding utterances, the coding decision was based upon the entire set of dependent utterances. For example: "You think it has to do with the poles? What do you mean by that?" Since the interpretation of the second utterance depends upon the first utterance, these two utterances were coded as if they were one.

First, all codable utterances were numbered. Second, decisions about the beginning and end of a test were made. Most tests begin with a prediction or selection code (Why do you pick these?). However, if the examiner asked the subject to test a particular pair of bottles or to demonstrate the generality of a finding ("Could you show me that width matters with another set of bottles?"), that request was included in the following test. Third, each numbered utterance received a code. Seven intervention types were seen as follows:

1. Standard Probes (S)

These were standardized questions used in an attempt to elicit comments from the subject concerning his strategy and reasoning processes. These probes occurred at three different times during a test: after rods or bottles are selected; before outcomes are observed, and after outcomes are observed. They were asked in a fixed sequence:

a. selection [after rods or bottles are selected]
"Why did you pick these?"
"Why did you pick these rods [bottles]?"

b. prediction [before outcomes are observed]
"What do you think will happen?"
"What do you think will happen now?"

c. conclusion [after outcomes are observed]
"What did you learn?"
"What did you learn from trying them?"
"What did you learn from trying these rods [bottles]?"
"What did you find out?"

Minor variations in phrasing were coded as S probes, but more radical variations were coded as NS (see type 2).

2. Non-Standard Probes (NS)

The non-standard probes were variations of the standard selection, prediction, and conclusion probes, which were attempts to derive esssentially the same information in a somewhat different manner. The NS probes were more specific than the standard ones, often placing an emphasis on, and drawing the subject's attention to, a specific experiment at hand.

a. prediction "Which one do you think will light up more?"

b. conclusion "But you found out that _____?"
"What did you find out about the light bulbs here?"
"So what did you see here? which one's brighter?"
"What does that tell you about brightness?"
"What is happening here?" [When it doesn't follow an unclear utterance.]

3. Clarification

These probe types followed unclear messages by the subject, and were an attempt by the examiner to clarify the subject's utterance. This code was used when it was quite clear that the examiner was trying to understand the meaning of the subject's previous utterance.

a. message/strategy "Which one would be lighter?"
"Which things?/The metal things?"
"What's almost the same?"
"When you say bigger, what do you mean by bigger?"
"So what do you think is happening here?"
[following unclear message by subject]

b. prediction "You think the WL will be brighter?"
"You think they'll be the same?"
"How do you think it will light up? Like this one?"

c. conclusion "Could you tell me that again? I'm not sure I followed you."
"Which one lit the most?"
"You think it has to do with poles? What do you mean by that?"
"The Z isn't any good? What do you mean by that?"
"You thought it was better because of what?"
"It seems to you the main thing would be electrode length?"

4. Justification

This type of probe followed either a prediction or conclusion by the subject. The examiner then asked for additional evidence or explanation to substantiate the subject's statement.

a. prediction "Why do you think that [will be brighter]?"
"Why do you think it should light up more?"
"Why do you think that might be true?"
"Because?" [following prediction]
"Why?" [following prediction]

b. conclusion "Why don't you think that the amount of fluid makes a difference?"
"How do you know that?"
"Why do you think that's true?"
"Why do you think this bottle and that chemical may be lighting this one brighter?"

5. Challenging understanding and goal structuring

These prompts were of two types.

Type 1: Here the examiner was questioning the subject's knowledge of the effects of the variables and the types of conclusions that may be drawn in a given situation. Often this was done in the form of a counter-example or opposing view designed to test the strength of the subject's convictions.

Examples: "Can this tell you for sure what's causing it?"
"Somebody said 'bottle size matters'. Do you think they're right?"
"What if the pole was up more, which do you think would light less?"
"Do you think that's true for all bottles?"
"Would that be a good way to figure it out?"

Type 2: This type of prompt was in the form of suggestions or questions by the examiner which attempted to keep the subject "on track." Often several of this type of intervention were found in a row as the examiner seemed to be attempting to guide the subject's thinking in a desired direction. Only the first utterance in a series of this type was coded.

Examples: "I asked you about the amount of fluid."
"Why don't you look at those?"
"What have you figured out so far?"
"Let's look at this. Maybe something else is happening here."
"What are you looking for?"

6. Reassurance

Any remarks made by the examiner designed to encourage the subject to continue or to provide nonspecific feedback concerning the reasonableness of the subject's activities.

Examples: "That makes sense."
"Very good."
"It's whatever you think."

7. Encouraging Self-monitoring

This refers to instances in which the examiner asked the subject for reflections on the subject's predictions, conclusions, or activities. It was also used when the examiner indicated that the subject may have made a mis-statement or misperception.

Examples: "Are you sure?"
"What about this? Do they still look the same?"
"Have you ever thought about writing it down?"
"I think this may be a little brighter. What do you think?"
"But you said you were going to test ['length']."

Two coders independently coded all four instructional sessions using the Examiner Intervention Code. Average interobserver agreement per session was 82% (range 74-91%).

Case Studies of Progress Across Sessions

Subject No. OP4

Quantitative Results. - OP4 was observed during six IV task sessions. Two of those sessions (the first and last) were devoted to the bending rods task and the remaining four were spent on the conductivity task (the first and last conductivity sessions were baseline observations and the second and third were instructional sessions). During the first session, the bending rods task was administered twice. Each of these sessions was coded using the procedures developed during the assessment phase of the project and seven of the same dependant variables were derived: generality and articulation of conclusions; use of explanatory principles; intentionality and explicitness of predictions; proportion of unconfounded tests before the summary; number of unconfounded

variables. In addition, the examiner intervention code was used for the two instructional sessions.

The results of this quantitative analysis are displayed in Table 70 and Figures 1-7. Because OP4 was the only subject who received two instructional sessions, his scores for these sessions were averaged in the tables and figures. Figure 1 illustrates that OP4 showed a substantial increase in the construction of unconfounded tests between the first two administrations of the bending rods task (from 0% to 50%). This increase was similar to but more dramatic than the increase shown by the average normal ninth grader during the assessment phase. The introduction of the new conductivity task in session 3 produced a sharp decline in unconfounded testing (to 18%). Gradually, over the course of the next three conductivity sessions, OP4's tendency to construct unconfounded tests again increased (to 42%). During the final bending rods session, OP4 exceeded his previous performance on the bending rods task during test construction phase 2 by constructing 69% of his tests in an unconfounded manner.

On another index of mastery of the IV strategy, the number of variables tested in an unconfounded manner, OP4 showed performance that was consistent with, but less variable than that measured by the proportion of unconfounded tests. During the first bending rods session, he went from no variables tested to two variables (out of 5 possible) tested in an unconfounded manner. On the conductivity task, he went from 1 to 2 variables (out of 3 possible) tested in an unconfounded manner. By the final bending rods session, OP4 tested all five variables in an unconfounded manner.

Additional indices of the IV strategy are available from the test choice and test evaluation phases of the bending rods task. Although these phases were administered only once, during the first session, they show that OP4 had not mastered the strategy by the end of that session. He made 3 out of 5 correct choices, but gave only 1 out of 5 correct explanations on test choice. On test evaluation, he made 3 out of 6 correct choices but gave only 2 out of 6 correct explanations. Thus, on all measures of the use of or preference for unconfounded tests, OP4 made substantial progress across the six sessions.

When the additional variables of task approach are examined, some patterns of performance which are consistent with that seen in unconfounded testing become apparent. For example, OP4 used appropriately specific conclusions (i.e., a level of generality similar to that of an average normal ninth grader, during the second and final administrations of the rods task. However, he tended to draw relatively general conclusions during all administrations of the conductivity task (See Figure 2). OP4 also showed high levels of intentionality and explicitness in his predictions during the second and final bending rods sessions and lower levels of these variables during conductivity (See Figures 5 and 6). Thus, on these three task approach variables and on both measures of unconfounded testing, OP4's performance on the second and final administrations of the rods task were comparable to each other and differed from his performance on the conductivity task.

However, the two remaining indices of task approach showed different patterns of performance. Unlike the normal ninth grade average, OP4's conclusions did not become more highly articulated between the first two administrations of the bending rods (See Figure 3). In addition, his conclusions showed a dip in

articulation at the first conductivity session. Over time, his conclusions became slightly better articulated, but they never exceeded his conclusions during the second administration of the rods task. OP4 showed an increase in the use of explanatory principles from the first to second administration of the bending rods (See Figure 4). By the second administration of the rods, he was using explanatory principles at a level similar to that of the normal ninth grade average. This tendency to use some explanatory principles was maintained during most of the conductivity sessions but disappeared during the final two session of both tasks.

A summary of the indirect instructional probes OP4 received is illustrated in Figure 7. The intervention used with OP4 did not differ very much from that used with the other two subjects. On several indices (standard probes, nonstandard probes, clarification probes, challenges and goal structuring and self-monitoring), he received approximately the same percentages of probes per test as did OP9. OP4 did receive substantially more justification probes than did OP9 (and slightly more than did OP5) and slightly more reassurance than did either of the other two intervention subjects.

Due to the absence of normative data across more than one or two sessions, it is difficult to evaluate OP4's performance on some of these dependent measures. Given the nature of the task, the index of unconfounded testing should show increases over time if the subject is attempting to improve his task performance. Also explicit and intentional predictions and highly articulated conclusions indicate an ability to express oneself in an appropriate way in this setting. However, it is less clear whether decreases in the generality of conclusions or in the use of explanatory principles constitute improvements.

What is most apparent from a number of these quantitative indices is that task familiarity and content are important factors in OP4's performance. On several indices (proportion of unconfounded tests, explicitness and intentionality of predictions, and specificity of conclusions), his performance on the second and final administrations of the bending rods task are more similar to each other than they are to his performance on the conductivity task. Thus, as he became more familiar with the rods task, his ability to contruct unconfounded tests and to make clear and planful predictions increased.

On some of these same indices (e.g., intentionality of predictions, unconfounded testing) he showed a similar pattern of steady increase over time in the conductivity task but at a lower level of performance. There is little or no evidence from these quantitative data that OP4's performance was affected in any specific way by the two sessions in which the examiner explicitly attempted to intervene.

Qualitative Results. - Inspection of the videotapes and preliminary coding sheet from both the first and final bending rods sessions with OP4 revealed striking performance differences. At the beginning of test construction phase 1, it was clear that OP4 did not prefer unconfounded testing. His spontaneous tests were either confounded or were compensation tests (i.e., tests where variables are intentionally confounded in order to make the two different rods bend equally). This tendency to construct confounded tests was fostered by his inappropriate rod selection strategy: After each rod was

tested once, it was put in a "used" pile and the next two rods were selected from those that remained. Thus, after several tests were made, the remaining "unused" rods were not conducive to any unconfounded tests. This selection strategy also affected his predictions. Initially, OP4 showed some prior intentions to test particular variables, "Well, I wanted to see if this plastic one would go as long as this wooden long one." After several tests, he was left with pairs of rods to test about which he had no ideas. His predictions for these tests showed a trial and error approach, "I don't know, I just picked any one that time."

He began to construct unconfounded tests when tests of particular variables were prompted by the examiner during test construction phase 1. He constructed two unconfounded tests when prompted to do so. One clear advantage of these tests for OP4 was that he did not use his immature, array-bound selection strategy. Thus, he began to show a growing awareness of the need to control all but one variable when asked to test that variable. The test choice questions show that his choice of unconfounded tests was more advanced than his ability to explain why they were better tests.

During test construction phase 2, OP4 continued to employ his array-bound rod selection strategy for his first five tests. Thus, his growing ability to construct unconfounded tests was hampered initially by an inadequate selection strategy. His verbalizations during this phase show an increasing awareness of the need to have prior intentions about his tests: Most tests begin with predictions such as, "I wanted to test . . ." or "I wanted to see if . . ." He also showed his preference for unconfounded testing by indicating the similarities as well as the differences between the rods he selected, "They're pretty much equal. If you put like a small weight on here and a big weight." His performance during the test evaluation portion of the task was similar to his test choice performance.

The most dramatic change in OP4's performance occurred during the final session with the bending rods, which took place after the four conductivity sessions. His array-bound selection strategy had disappeared, and his testing had become largely unconfounded. His predictions were highly explicit and showed a high degree of intentionality. He showed he was quite capable of testing all five variables in an unconfounded manner. All evidence of confusion, uncertainty, and hesitancy about the task disappeared. Although the source of some of these sophisticated behaviors will be found in the initial rods session, their mastery had to have occurred sometime during the intervening conductivity sessions.

In order to describe his behavior and that of the examiner during the conductivity sessions, a graduate student who was an experienced LD practitioner wrote up a specimen record for each session. The specimen record itself constitutes a fairly objective description of selected parts of the session. Contextual information, where necessary, is provided between parentheses. More evaluative comments about the description are bracketed or appear in the summary.

The chemicals used in the conductivity sessions with OP4 were labeled: session one: H, J, B; session two: X, V, Z; session three: W, L, Z; session four: J, X, M. In each case the first chemical was stronger than the second, the second was stronger than the third. The third was a nonconducting

chemical, and when combined with either one or both chemicals, it diluted them. Also, a combination of the first and second chemicals produced a weaker solution than the first chemical alone.

At the beginning of Session 1, OP4 was seated before an array of chemical bottles and a conductivity test box. The examiner entered and sat down. She pointed out that the bottles on the table contained chemicals H, J, and B. The purpose of the activity, she explained, was to determine what mattered for brightness. She explained that the chemicals in each bottle were lettered. OP4 stared at the bottles while the examiner spoke. The examiner proceeded by demonstrating with two of the bottles. The examiner called OP4's attention to the demonstration by asking him to "watch" as she traced (pointed) the path which the electricity took as it traveled to the lightbulb. During the demonstration the examiner enunciated the key vocabulary words (electrode, chemicals, brightness) clearly. OP4 chewed gum vigorously throughout the demonstration; he nodded in understanding. The examiner pointed out that one bottle "lights more" than the other one. When asked why this might happen, the student grinned, laughed slightly, and said that he had no idea. The examiner smiled reassuringly. OP4 then made eye contact with the examiner and attempted to explain the reason for brightness -- "maybe more/less chemicals." The examiner suggested that might be true, but more things may matter for brightness. OP4 laughed uncomfortably; the examiner ignored the reaction. She then offered him paper to keep a record of his findings as he proceeded with the testing. He did not use the paper.

With one arm resting on the chair armrest, OP4 chose just one bottle to test, JB (a small bottle, the chemical was diluted with the nonconducting chemical; short electrodes). His reason for choosing this bottle was because he " . . . just wanted to test the small 'ones' to see how they work." The examiner did not understand what the child had said and leaned forward to indicate she was not sure of his response. He repeated, "I was just testing the smaller ones." He concluded that " . . . smaller ones put a little bit more out." OP4 removed the bottle from the box and placed it behind those he had not tested. The next test was another single bottle test - a large bottle with long electrodes. He " . . . picked a bigger JB one, and ah, seems like there's more power going to the big rods, to light up more." He appeared to be referring to the electrodes or to bottles (bottles may be more salient), although he did not specify this in his speech.

He then decided to choose one (HJB) bottle. Glancing at the array of bottles he mumbled, " . . . just testing the, you know all the, the numbers, I mean the letters and stuff." [His lack of an understanding of the precise vocabulary to use seems to impede his formulation of ideas.] The examiner interrupted and asked, "What do you think is going to happen?" He responded, "I think it's going to go on, a lot." Before continuing with the test he hesitated and added, "A real little, either a lot or a little." [He appeared to be guessing, perhaps in hopes of satisfying the examiner.] He concluded it lit up "real little." He then chose H (the strongest chemical) for " . . . a change . . . don't know, I just picked it." He thought that it would, " . . . light up pretty, pretty much. Hopefully." (He laughed.) When questioned by the examiner he concluded that, " . . . the H one lights up more." He pushed the control switch one more time [in satisfaction] and removed the bottle placing it behind the bottles not yet tested. He slowly proceeded to the next test. He chose a small bottle H with a short electrode, " . . . to see if the

bigger one's more than the little one." He concluded that they were about the same. After considering his response he added, "Can you put like two in at the same time?" The examiner crisply responded, "Yep." OP4 searched among the bottles he had already tested, snapped his fingers, made a selection, and proceeded [with renewed interest]. Apparently satisfied with this new insight of comparing two bottles, he employed both hands while setting up this experiment. In an even tone the examiner asked, "Okay, now why'd you pick those?" He had chosen H (the strongest chemical) and an H with a short electrode. Both bulbs lit up, he immediately glanced at the examiner. The smile on his face broke the look of boredom as he said, "They both light up real light." He watched the examiner as she recorded the results. He then went back to the use of one hand with the other arm remaining on the armrest.

The fifteenth test of this session was significant in that OP4 again reverted to a single bottle test. [After it earlier appeared that he understood that comparison was important while considering brightness.] He chose the nonconducting chemical and short electrodes. His reason for the choice was, "To um, just see the difference." [Note, he did not specify the nature of the difference he was going to observe.] The examiner asked him for his prediction of what would happen. "The light will go on real light. I mean, not real light; really low." His last compromising remark suggested a search for an understanding as to what was happening, but he was unable to comprehend the results of the previous test and could not apply previous test results to his current testing. The bulb did not light; OP4 continued to push the control switch [as if the problem could be in the switch] and picked up the light shade while examining the bulb closely. OP4 sighed and mumbled, "It didn't go on at all." He looked closely at the bottle before removing it from the box.

At the conclusion of session 1, the examiner selected the bottles used in the demonstration and asked, "Do you remember how they lighted up before?" OP4 was not able to articulate which variable might have made a difference in that test. He shook his head while gazing at the conductivity box and was at a loss as to what to say about brightness. A series of questions from the examiner enabled OP4 to come to the conclusion that, "B ain't no good." Notice that he did not formulate generalizations relating these variables to the cause of degrees of brightness.

Examiner: You found out some things. What . . .
OP4: Well, this one's smaller and it don't take as much.
Examiner: Yeah. Okay. Anything else?
OP4: And the, the length and the ah, iron things that go down there.
Examiner: So that has something to do with brightness?
OP4: Um-huh.
Examiner: Okay. Anything else? What'd you find out about the chemical?
OP4: Just the H,J, and B together ain't that good.
Examiner: And what about this one?
OP4: The H is good. (It was the strongest chemical.)
Examiner: What about the other chemicals?
OP4: The B is not good (it was the nonconducting chemical) and I never found out what the J was.

Although the subject did conclude that B did not conduct, he did not observe its effect on the other chemicals, nor did he articulate that the strength of the chemical was a variable in "what mattered for brightness."

One week later, an intervention session (Session 2) occurred during which the examiner was permitted to ask probing questions of specific types.

The examiner explained that she had brought different chemicals - X, V, Z. She proceeded with a demonstration and concluded the test by asking OP4 if he had any ideas as to what mattered for brightness. He shrugged his shoulders and responded with a mumble indicating that he did not know. The examiner repeated the question, and OP4 then suggested that maybe the (smaller) size mattered. The examiner reminded him that he must note all the things that make a difference. OP4 surveyed the bottles. The examiner offered the subject a pencil and paper on which to take notes. He ignored the suggestion. [He appeared to be more confident than in session one perhaps because the examiner was more (verbally) involved in this session. Throughout the testing the terminology he used to express the variables was vague. For example, he chose one bottle because of the "stem things." Regarding bottle choice, he wanted to "see what these different 'ones' are in here."]

After having concluded that X (the strongest chemical) with "longer things makes it go a lot," OP4 was encouraged by the examiner to compare large bottle X with a small bottle X with short electrodes. The examiner questioned, "Okay, now what do you think might happen with these two?" "They both light up." OP4 pointed to the one (large bottle) he thought would light more. "Why?" "Because more power is going through longer things." With encouragement from the examiner he tested the bottles and concluded the "bigger" one produced more light. The examiner asked for a clarification of the term "bigger." He said he meant the bigger bottle and "metal wire." To challenge his understanding, the examiner pointed to the bottles and questioned, " . . . looking at these two bottles could you tell for sure whether it was the longer metal things or the bigger bottle?" After gazing intently at the bottles, OP4 responded, "Probably both. I don't know." Urging him to evaluate his response the examiner asked, "Is there some way you could tell using those other bottles, whether it was the long things or the big bottles?" After looking at the array of bottles, he hesitated, picked up one bottle, then put it back. The examiner urged him to find others that might prove the point. He chose two bottles (X,XV) the same size, both with long electrodes. [He seemed to be unsure as to how to control for variables.] He pointed to the bottle he thought would light more. The examiner asked why he had chosen that one. He replied, "Because it's got the, the V in it," and added, "The V ain't that good." [Though this was an unconfounded test for chemical strength, one is not sure whether in his mind he was testing strength or bottle size.] The examiner asked what did he find out? "They're actually more powerful." [A rather ambiguous response. Even with probing questions, the subject seemed unable to isolate, test, and articulate which variables are related to brightness.]

During the session the examiner asked OP4 to review what he had learned about brightness. He paused. "X, good; V, all right; Z ain't no good. The examiner asked him to clarify what he meant by "no good." He responded by saying the Z was no good because it doesn't light up. His eyes wandered about the room. The examiner then asked, "What did we [by using thi pronoun she

made herself a partner in the activity] find out about the other things?" He mumbled, "'Bout the stems?--It doesn't matter." Challenging the subject the examiner asked, "Do you know for sure?" After hesitating he chose small X with short electrodes because it "lights up" and V was chosen because "of the longer things." He thought they would be the same brightness. After some hesitation he stated, "V is good with long things and big bottle X is better--it doesn't matter about the things." The examiner set up an experiment to guide his thinking. She selected two small VZ bottles, one with short electrodes. OP4 agreed it might be a good test. He then chose two bottles with "different stems," and predicted the "little ones" would light up more. However, he concluded, "The long ones are better."

Going back to the original test, OP4 reviewed, "The Z isn't that good, V is all right, X is all right." The examiner continued, "Is there anything else about these two bottles that might be equal?" He suggested the 'long ones" are better. The examiner asked, "Okay. Is there anything else about these two bottles that are different that might have something to do with brightness?" The subject responded with a question, "How big they are?" [This last response indicated he was still having difficulty integrating the isolated bits of knowledge he had gained to come to any generalizations which would explain all the things that mattered for brightness. The examiner's questions and the inclusion of herself ("What did we find out?") seemed to positively affect his attitude toward the task in that he continued attempting to give answers to her questions.]

During the Third Session OP4's performance fluctuated. At times he appeared involved in performing comparison tests, but seemed unable to maintain a meaningful direction in the selection of tests.

OP4 picked W (the strongest chemical) because it seemed "more better," and he knew Z (the nonconducting chemical) "isn't that good." He wanted to see if W was medium and concluded that it was indeed medium. "How did you know it was medium?" the examiner questioned. [Perhaps thinking the examiner was indicating his answer was not correct] he added, "Well, it's not really medium, but it's almost." The examiner asked him if there was a medium one to which he could compare it. He then proceeded to compare ZW (W diluted with nonconducting Z) with L (the weaker chemical) concluding that they were about the same. Although he explained the "same" referred to "lightness," he appeared to be puzzled by the conclusion.

[OP4 became more involved in the task.] He decided to test a small WL bottle with a large ZW because "they got the two more strongest chemicals in them." He found "it" lights up more. He concluded W and L were best for brightness. The examiner asked, "Better than . . .?" He immediately responded that he wanted to test "that" and proceeded to compare W (strongest) with WL. He thought the WL would be brighter because the L would make it "more powerful." After testing he concluded (correctly) that L took power away from W. He was puzzled. [He appeared unable to adjust his thinking when contradictions to his predictions occurred.]

The examiner attempted to direct his attention to the electrodes. OP4 examined the bottles closely, but was not convinced that electrodes mattered. He rested his face on his hands, leaned back in his chair and folded his hands. The examiner called his attention to the top of the bottles. "Do they

(electrodes) look any different when you look at them on the top?" "No," he answered firmly. She then indicated width with her fingers. He was still puzzled as to why that would make a difference. OP4 glanced to the side [perhaps at a clock]. He examined the bottles, but he could not determine which bottles to choose to test for the effect of electrode distance. He finally agreed to test a small bottle, with a large bottle LZ (closer electrodes). He thought they would be about the same. After testing he concluded LZ (closer electrodes) was "better." He hesitantly suggested that the reason it was better was because "those things" were closer. The examiner reinforced the concept, " . . . so you think maybe the stems might have something to do with it? Are there some other bottles you could look at for the stems?" He seemed puzzled, so the examiner suggested two small bottles with chemicals ZL; the only variable that differed was electrode distance. He hypothesized and concluded that when the "things" are closer together it "makes it better." During this session the examiner used the subject's vague terms ("those things", "stems")

[Although OP4 had discovered the significance of some of the variables (strength of the chemicals and width of the electrodes) he did not use this information to make a generalized statement regarding brightness.]

The Fourth Session was an observation session during which the examiner was again scripted. As she entered the room, OP4 was leaning back in his chair, relaxed. The examiner sat down, pointed out the chemicals she had brought --X, J, and M. She performed a demonstration test, gestured to the array of bottles and asked if he had any guesses (as to what might happen). The subject immediately pointed to one of the bottles. When asked why he had chosen that one, he responded, "just guessed." He refused the offer to use a pencil and paper to record his test results.

[OP4 seemed to sense the examiner was more reserved during this session in that he was not as responsive as in the previous session. Many of his answers were mumbled; he used vague terminology in his responses.]

"Why'd you pick those?" (XJ, short electrodes; XJ long electrodes) The subject responded, "They got different 'things' on 'em. They got, you know, stem things on there." He pointed to the one he thought would "light more." He concluded the short electrodes "don't light as much as the one with the long things on it." He then chose a large bottle, J, and a small bottle J, with short electrodes to "see how the bottles are." He concluded the smaller one lights more than the larger. His next choice was to "test out the bottles." (MX, large; XM, small) Although he thought the smaller one would be brighter, he concluded that the "big one was 'better.'" He proceeded to test MX and XM just to see "which one would go brighter." He only had a "feeling." He concluded they were about even.

He moved from testing bottle size to testing whether the length of the "thing-stems" mattered. He used two small XM bottles, one with short electrodes. He concluded electrode length was relevant. [His terminology continued to be vague, but his tests were largely unconfounded or confounded only by electrode length. He also systematically tested each important variable in turn.]

At the conclusion of the half-hour session the examiner returned to the original test of the day. OP4 did not recall the results. He thought one was bright and had a J in it. M, he thought, did not work (correct assumption) and X worked. The examiner asked if there was anything else that mattered for brightness. OP4 responded, "Things farther apart." The examiner asked if bottle size mattered. He said he thought it mattered a little bit. The examiner asked, "What about anything else?" OP4 stifled a yawn. The examiner suggested the possibility of considering something else. She picked up a bottle which the subject looked at and responded, "I guess it matters . . . the length of these." "What happens with the length?" questioned the examiner. OP4 did not remember. The examiner encouraged him to respond by interjecting quickly, "Wait, wait. What do you think? Do you have any guesses - - if you had to guess, which one is brighter." OP4 yawned and pointed to one.

In Summary, during the first session the responses of OP4 consisted of ambiguous phrases and abbreviated sentences. ("Just picked it." " . . . just testing.") Although he appeared to be selecting the bottles carefully, and although he seemed to place them in an organized position after using them, he did not use this organization to guide his hypotheses and conclusions in future tests. This approach to the task may indicate that he did not comprehend that learning is cumulative and that he might need to make comparisons in order to achieve a complete understanding concerning all the factors that influence brightness in the conductivity task.

Probing questions from the examiner during the intervention session did elicit more involvement in the activity in that he was willing to perform further tests to verify results. Although many correct conclusions were drawn, he was not able to stabilize them for later use. For example, he tested for electrode length, electrode width, and chemical strength. However, he did not use this information when he was asked to summarize the test results.

The contrast in verbal responsiveness between the observation and intervention sessions seems to imply that adult interaction may be a source of motivation required for learning.

Subject No. OP5

Quantitative Results. - OP5 was observed for six IV task sessions. In the first and last sessions the bending rods task was administered and in the remaining four sessions the conductivity task was given. However, during the first baseline conductivity session, the chemicals in the vials were improperly prepared which meant that the feedback they produced was inconsistent. Therefore, an additional baseline session was administered (session 3) and only one intervention session was held (session 4). Data from that initial conductivity session were not analyzed.

The results are displayed in Table 65 and Figures 1-7. Figure 1 illustrates that the proportion of unconfounded tests produced by OP5 across the five sessions fluctuated from a low of 13% (TC2) to a high of 64% (Intervention). With one exception (the final conductivity session), he produced a greater proportion of unconfounded tests on the conductivity task than he did with the bending rods. (Note: OP5 conducted only three spontaneous tests in this session.) This is a reversed pattern from the one seen with OP4. However,

OP5's performance profile seems somewhat similar to that of OP9 on this measure. Both OP5 and OP9 produced a smaller proportion of unconfounded tests in TC2 than in TC1 and both showed a dramatic increase in unconfounded testing when the conductivity task was first introduced. This pattern of results was quite unlike that of the ninth or fourth grade controls.

On another assessment of mastery of the IV strategy, the number of variables tested in an unconfounded manner, OP5 showed more consistent results. He tested 1 out of 5 variables in an unconfounded manner in TC1, 2 out of 5 in TC2, and 5 out of 5 in the final rods session. During the three conductivity sessions he tested 1 out of 3 variables in an unconfounded manner in the first, 2 out of 3 variables in the second, and 0 out of 3 variables in the third. With the exception of the final conductivity session, OP5 showed a steady increase, over time, in his ability to test each important variable in an unconfounded manner in both task settings.

Two additional indices of the preference for unconfounded tests also indicate a steady increase in mastery of the IV strategy. During the test choice phase of the initial rods session, OP5 made 3 out of 5 correct choices and 1 out of 5 correct explanations. By the test evaluation phase, his performance had improved considerably. There he made 5 out of 6 correct choices and 3 out of 6 correct explanations.

Therefore, when all measures of unconfounded testing are included, it appears that the relatively low proportion of spontaneous unconfounded tests that OP5 produced during most of his sessions may be an underestimate of his mastery of the IV strategy. Several factors seem responsible for this underestimate. First, in several sessions, OP5 produced very few spontaneous tests. Second, he tended to produce most of his unconfounded tests in TC2 and in the final rods session in response to prompts for particular tests by the examiner. Thus, the proportion of total tests that were unconfounded would be much higher than the proportion of unconfounded spontaneous tests reported in Table 65 and Figure 1. Third, several of his tests in the initial rods session were either compensation tests or were tests of nonstandard or inaccurately perceived variables. (For example, he spent a lot of time investigating the effect of the two holes in the stand in which the rods were inserted.)

When the additional variables of task approach are examined, a slightly different picture of performance gains over time emerges. Figure 2 shows that with the exception of TC2, OP5 gave increasingly general conclusions over time. Like the normal ninth graders but at a much higher level of generality, OP5's conclusions increased in specificity from TC1 to TC2. This same pattern was true for the other two remediation subjects as well. However, after TC2, OP5 tended to draw extremely general conclusions from his tests. (For example, "So that proves the theory that weight does make a difference.")

The pattern of articulation codes is the inverse of the specificity codes: Over time, his conclusions became more highly articulated and less specific. TC2 is again the only exception to this pattern. OP5 and OP9 show a similar pattern of performance between TC1 and TC2, but afterwards their curves diverge with OP9 decreasing in articulation to the level of OP4. None of the three remediation subjects performed like the average of the ninth or fourth grade control groups during the rods pre-test session.

OP5 is also extreme in his use of explanatory principles. (See Figure 4.) Although he and OP9 are alike on this code (especially during the conductivity sessions), they both use many more explanatory principles than OP4 or the normal ninth and fourth grade groups.

A somewhat different pattern emerges with respect to OP5's prediction data. His predictions are much more intentional in the conductivity task setting than in the rods task. This pattern is almost the reverse of OP4's performance on this variable (with the exception of TC1). It is unlike that of either OP9 or either of the normal control groups. However, his predictions vary in explicitness over time, first increasing (like both normal control groups and OP4) between TC1 and TC2, then decreasing, then increasing again. Between TC2 and the final conductivity session, OP5's performance is identical to that of OP9. In addition, both subjects' predictions increase in explicitness between the final conductivity session and the final rods session.

A summary of the indirect prompts used with OP5 is presented in Figure 7. OP5 received a somewhat different proportion of prompts in several categories than did the other two intervention subjects. He received many fewer standard prompts and somewhat more nonstandard probes. That was due to two factors. First, he tended to supply, without prompting, a continuous narration of his activity, thus making additional prompts from the examiner superfluous. Second, his overly general style of speaking about tests caused the examiner to direct her inquiry to the particular test in front of him using nonstandard probes such as, "What did you find out about the light bulbs here?"

OP5 also received more prompts focused on self-monitoring and somewhat more challenges and goal structuring probes. Self-monitoring prompts were used in order to get him to notice discrepancies between his predictions and conclusions or to pay closer attention to particular vial attributes or outcomes. Challenges to his understanding were used to get him to question the validity of his conclusions and to appreciate the explanatory power of unconfounded testing. Finally, OP5 received fewer requests for clarification than did the other two subjects. This may be due to the fact that his lengthy narratives provided sufficient information about his thinking.

Qualitative Results. - What these quantitative results indicate but cannot fully describe is the verbose style of OP5. Over time, his ability to construct unconfounded tests increased in a slow, but steady fashion. However, his predictions and conclusions did not seem to be in synchrony with this gradual mastery of the IV strategy. Often he would make highly general, idiosyncratic explanations that were only superficially linked to the particular rods or vials that he had selected to test. For example, in the baseline conductivity session, he produced an unconfounded comparison of two chemicals by comparing mixtures XV and ZV. The two vials were identical in size and electrode length. The transcript from this part of the session follows. (OP5 has just observed the outcome of this test.)

E: So what'd you find out?

OP5: I think V, well, X and V are very, seem to be powerful chemicals, but ah, it's like taking X is a positive, V is a positive and you mix it together and you get more positive. Whereas, V is a positive and Z is

a negative and then you drop down. You may still have a positive charge, but . . . you have less of a charge. So, it would be the type of chemical, the amount, ah size of electrodes . . um, what else can it be? There's the size of electrodes, uh, certain mixtures of chemicals . . .

E: Which ones?

OP5: Well, I think, X and V are both bright, but ah, Z is pretty dull which we were able to prove with this one here. When we put this in, it didn't bubble at the bottom here, but ah, outside of that, it didn't even light up. So it's sorta like a negative, you could consider it. And, V and uh X are positive.

As you can see, OP5 seemed to feel that each test was an opportunity for him to expound in detail about the task as a whole. His actual observations about the outcome of that particular test were embedded in quite a bit of unnecessary verbiage. The ability to "go beyond the information given" is often seen as a conceptual strength. However, solid scientific theories do depend upon an adequate data base. Because OP5 paid too little attention to the attributes of his data and to the outcomes of his tests, he often misinterpreted his own experiments. He tended to distort or ignore data that disconfirmed his theories instead of modifying his inadequate theories to better fit the data. Although the above excerpt was selected to illustrate a point about OP5's performance in these sessions, it was not an unusual sample of behavior. He repeatedly summarized all previous conclusions throughout the session and often hypothesized elaborate explanatory principles to support his conclusions.

Overall, OP5 gave a confident impression. His prosody, expression, and posture suggested control and assuredness. When delivering a summary of his findings, he sat back, relaxed, and spoke clearly, using bottles as illustrators or pointing with his pencil. Superficially one would conclude OP5 has an excellent vocabulary. He uses many abstract words (e.g., neutralize, equivalent, resistance) and tries to explain differential conductivity using constructs such as negative and positive. However, these abstract terms and lengthy explanations may mask rather than reveal his understanding of the phenomena. When pressed to explain his ideas more fully, OP5 often ended up confusing both himself and the examiner. OP5 seldom responded directly to a question which required a definite answer, "Are they the same or different? Which one will be brighter?" Frequently, his answers were qualified with "maybe, probably, not sure, but."

OP5's verbal reasoning was most coherent when he made written notes during the second conductivity session. He referred to his notes when asked to summarize his findings and when he needed to remember his previous tests. He also used the notes to help him select tests. When he used his notes, his summaries were concise and consistent, although still qualified. Unfortunately, OP5 did not continue to make or to use written notes in his subsequent sessions.

Several instructional strategies might prove useful with a subject like OP5. First, his ability to use his own written notes to organize his thoughts and record his findings could be encouraged. Perhaps he could be instructed to monitor and correct his written verbal expression first and then to transfer

these metalinguistic skills to his oral expression. Second, he needs to be persuaded to use data from his tests more effectively. His hypotheses and inferences need to be based upon data and revised when they are not supported by the data. Third, he should be encouraged to define the abstract terms he uses and discouraged from making elaborate and idiosyncratic analogies like the one he made between differential conductivity and positive and negative numbers. These abstract concepts and explanations by way of analogy could be useful if they were grounded in a clear understanding of the phenomena under investigation. However, when OP5 used them they seemed to distract and confuse him.

Subject No. OP9

Quantitative Results. - OP9 was observed during five IV task sessions: two bending rods sessions (the first and fifth) and three conductivity sessions (the second, third and fourth). The instructional session occurred during the third conductivity session.

The results of this quantitative analysis are displayed in Table 65 and Figures 1-7. Figure 1 illustrates that OP9 showed a slight decline in the proportion of unconfounded tests from test construction phase one to two (from 22 to 14%). This decline was quite different from the increase in unconfounded testing from TC1 to TC2 for both the average ninth and fourth grade controls in the assessment phase. The primary reason for the decrease was an increase in the proportion of compensation tests (tests where variables are intentionally confounded in order to get the two rods to bend the same) between the two task administrations. His proportion of unconfounded tests increased (to 60%) during the first conductivity session, but then declined to 44% and 33% by the final two conductivity sessions. Again, this decline may have been due to a conscious decision on the part of OP9 to choose confounded over unconfounded tests in order to make cross-test comparisons. (See Qualitative Results below.) OP9's proportion of unconfounded tests increased dramatically during the final administration of the rods to 60%. This was the same degree of unconfounded testing as he had demonstrated during the first conductivity session. As high as this proportion appears to be, it may be an underestimate of his mastery of the IV strategy, since his remaining tests during this session were all compensation tests.

On another assessment of the IV strategy, the number of variables tested in an unconfounded manner, OP9's performance was seriously affected by his preference for compensation and confounded tests. During the first bending rods session, he went from two variables tested to zero variables tested (out of 5 possible) in an unconfounded manner. During the three conductivity sessions, he never tested more than one of the three possible variables in a consistently unconfounded manner. However, his mastery of the IV strategy became apparent during his final rods session when, suddenly, he demonstrated that he could test all five variables in an unconfounded fashion.

Additional indices of the IV strategy are available from the test choice and test evaluation phases of the bending rods task. These two phases, administered during the first session only, demonstrated that OP9 showed a preference for unconfounded testing that was obscured when he constructed tests on his own. He made 4 out of 5 correct choices and 4 out of 5 correct explanations during test choice and 4 out of 6 correct choices and 4 out of 6

correct explanations during test evaluation. Thus, his appreciation of the superiority of unconfounded tests was shown most clearly on the more structured phases of the initial administration of the rods. When he was allowed to construct tests on his own, he was more likely to use alternative testing procedures (e.g., intentionally confounded compensation tests).

When the additional indices of task approach were examined, a different set of findings emerged. With respect to OP9's use of general conclusions, Figure 2 shows that he used less specific conclusions during the rods pre-test session than did either of the normal control groups (N9 and N4). However, his change during that session from more to less general conclusions paralleled that of the average N9 subjects as well as that of the other two intervention subjects. During the first two conductivity sessions, OP9's conclusions became more general, but by the final conductivity session, they increased in specificity. Finally, the last rods session found him making highly general conclusions. Thus, OP9 showed great variability in his use of general conclusions across sessions. This variability did not seem to be related to task characteristics or practice effects.

As early as session one, OP9 used highly articulated conclusions. (See Figure 3.). This degree of articulation was much higher than that used by the average 9th and 4th grade subjects, but was similar to that used by OP5. Both OP9 and OP5 declined in the degree of articulation from TC1 to TC2, unlike the normal control groups. The introduction of the conductivity task reduced OP9's performance further to a level similar to that of OP4. The similarity between OP9 and OP4 persisted throughout the conductivity sessions on this variable. However, OP9 was able to increase his proportion of highly articulated conclusions up to that of his TC2 level (75%) during the final rods session.

OP9 was unlike OP4 and similar to OP5 in his frequent use of explanatory principles. (See Figure 4.) For example, after observing the outcome of a conductivity test, OP9 explained what he saw in this way: "Because this (the electrode) goes down farther into it, and it can pick up more particles. And this can't pick up as many. But since J must be really strong, it can pass through." Both OP9 and OP5 were more likely to use these explanatory principles over time and in both task settings. In addition, they used them to a much higher degree than did either OP4 or the two normal control groups.

Although OP9's predictions were highly explicit and showed a high degree of intentionality during the first session with both the rods and conductivity tasks, this explicitness and intentionality declined over time. (See Figures 5 and 6.) This may have been due to his increased interest in developing explanatory principles about the variables. For example, in the rods post-test session he selected two rods that differed only in material. However, this difference was not marked in his prediction, "I picked these 'cause they're about the same length . . . and ah, I wanted to see how much flimsier this is compared to this." Thus, while he seemed to be aware of the variable he was testing (he marked it in his conclusion), he did not mention it in his prediction. This situation occurred frequently during the test sessions. Without normative data on the kinds of predictions subjects use over more than a single session with an IV task, it is difficult to evaluate the significance of OP9's apparent performance decrement. Our normative data on a single rods session showed that the average ninth and fourth grader

increased in explicitness between TC1 and TC2, but decreased in intentionality during this same time frame.

A summary of the indirect instructional probes OP9 received is presented in Figure 7. OP9 seemed to receive the same kind of intervention as did the other two remediation subjects. The one type of probe that he received much less frequently than did the other two subjects was the justification probe. In most other categories, he received approximately the same proportion of probes of that type as OP4.

In summary, the results of our quantitative measures of OP9's performance were complicated by his apparent interest in exploring issues related to but different from the IV strategy itself. During his initial performance on the rods, he was able to demonstrate his understanding of the value of unconfounded testing during the more structured phases of the task. This mastery was disguised during the two test construction periods by his interest in compensation tests. In addition, his conclusions were highly articulated and his predictions were quite explicit and intentional during the first session. When the conductivity task was introduced, OP9 showed an increased ability to construct unconfounded tests. Although his conclusions were no longer as well articulated during this second session, his predictions continued to show a high degree of explicitness and intentionality. Throughout most of the intervention period, OP9 found the IV task setting to be a rich source of ideas about flexibility or conductivity. These ideas were expressed in his frequent attempts to derive explanatory principles. By the final rods session, his mastery of the IV strategy was quite apparent in the proportion of unconfounded tests and, especially, in the number of variables tested in an unconfounded manner. Thus, these results indicate that OP9 found the IV task setting to be a stimulating environment for reasoning and problem-solving activity. The instructional intervention that occurred during a single session did not seem to affect his behavior in any observable way. However, his progress over time in mastery of the IV strategy indicates that the physical environment itself provided him with the kinds of experiences necessary for the consolidation of that strategy.

Qualitative Results. - In order to provide a richer picture of OP9's behavior, a graduate student wrote up a specimen record for each of his three conductivity sessions. Contextual information is provided between parentheses. More evaluative comments about the description are bracketed or appear in the summary.

The chemicals used in OP 9's conductivity tasks were labeled as follows:

session one: P,K,L
session two: J,X,M
session three: D,F,L.

In each case the first chemical was stronger than the second, which in turn was stronger than the third. Like the chemicals used with the other subjects, the third chemical was nonconducting and, when combined with either of the other chemicals, diluted them. Similarly, a combination of the first and second chemicals produced a weaker solution than the first chemical alone.

At the beginning of Session One, OP9 was seated at the experiment table, staring at the array of chemical bottles placed with the conductivity test box before him. As the examiner entered and sat down, he looked up over the conductivity box but did not look directly at the examiner. The examiner picked up a bottle she had set aside and explained that it had three chemicals in it--P, K, L. As she spoke, OP9 stared at the bottle. She then went on to demonstrate the procedure with this bottle and to explain what was happening. The subject watched this demonstration and explanation without ever taking his eyes off the equipment to look at the examiner. The examiner then took a second bottle, placed it in the box, attaching the appropriate wires, and pressed the button. She continued to hold the button down as OP9 looked at the setup. OP9 looked at the lights for several seconds, looked down at the chemicals and wires, and then glanced up at the lights once more as the examiner noted that one was brighter than the other. The subject nodded slightly, following which the examiner asked why he thought that happened. Continuing to look at the setup, the subject responded that there was a "certain kind of acid" in one and that the other had "other things added" that caused it to light less. The subject pointed to each bottle in turn as he talked about it. During this explanation, the examiner rested her chin on her hand, looking at the setup as the subject explained. Several times she nodded her head in agreement. When he had finished, she replied that he might or might not have been right and that she wanted him to figure out everything that would make one brighter and convince her of this. She remarked that it might help to take notes and gave him paper and a pencil should he decide to do so. He did not.

Upon completing his first test, OP9 removed one of the bottles and placed it behind the array on the table. Still looking at the array, he reached to remove the second bottle, KL in a big bottle with long electrodes, stopped, glanced once more from the setup to the array and then selected PK in a large bottle with short electrodes, leaving KL in the box. In response to the examiner's query, he replied that he picked them to see if K had "anything to do with it." As the examiner bent over to make a notation in her records, the subject added that he wanted to see if K would "take away most of the power of the P" and that he wanted to see the effect of the "shorter thing." Following the test, he concluded "that it's either that K takes away the power or the shorter thing." As in the demonstration and previous test, the subject pointed to the bottles while talking, but did not look directly at the examiner, even when she was looking directly at him.. [In his conclusion, the subject seems to demonstrate some awareness of confounded tests and the types of conclusions that can be drawn from them.]

Following the third test, OP9 reached into the box and switched the positions of the two large bottles, K with long electrodes and L with short ones. He did this without hesitation and without looking at the array of other bottles to his left. When he had finished attaching the bottles in the box, the examiner asked why he had switched them. OP9 replied that they each had "different outlets . . . back there and (he) wanted to see if that had anything to do with it." He then closed the box and pressed the switch, leaning forward as he looked several times from the lights to the bottles and back again. When the examiner asked what he had found, he released the switch and sat back in his chair. Still looking at the setup, he replied that the outlets made no difference.

Tests seven, eight and nine formed a series of unconfounded tests for chemical. On his eighth test, OP9's hand hovered over the bottles as he scanned them and made his selection. After several seconds, he selected K and P, both in large bottles with long electrodes. In response to the examiner's questioning, he replied that he picked them to see which would light more. He predicted that P would light more because "it has more acid in there." As in previous tests, he leaned forward as he pressed the switch and looked at the lights. When asked what he had found out, he sat back in his chair before replying that P "is the purer acid" while K "has additives in it." He again pointed to the bottles as he spoke to make it clear which bottle he was referring to.

Following the tenth test, the examiner asked OP9 if he knew everything that mattered for brightness, to which he replied that he did. The examiner placed the original bottles in the box, reviewed the results, and asked what things OP9 thought might have made a difference. The subject mentioned that in one of the bottles, the chemical was pure and caused it to light more. He also mentioned that one of the bottles had a shorter stem. He also hypothesized about the effect of L on the other chemicals: "But the L must be a harder substance to get through. It must surround the (stem), not very many particles of the acid get through." When questioned about his ideas of the effect of electrode length on brightness, OP9 explained that "the longer one would light up more because more of it (the chemical) gets onto this (the stem)." He also stated that stem length had no effect on the pure chemical. The examiner then pointed out that bottle size might also matter. OP9 affirmed this with a slight nod and a mumbled "right." The examiner remarked that she knew he'd noticed and asked him to show how one could tell whether or not bottle size made a difference. Without waiting for the examiner to complete the sentence, OP9 selected a large and a small bottle, each with long electrodes containing KL. As he connected the bottles in the box, he glanced back once more to the array, [as if to make sure he had selected the correct bottles.] When asked about his selection, he replied without hesitation that he had picked them because "they're both the same substance, and they each had a long stem" as well as having different bottle sizes. He hesitated slightly before forming a prediction and continued to gaze at the conductivity box. He then answered, "I think this one (bigger bottle) will light up a little bit more because it has more K to get through the L; but it's equal. And these are equal so . . ." As the subject paused, the examiner glanced up from the box and looked directly at him. After several seconds, OP9 completed his prediction and stated, "It should equal out." [Even though it was an unconfounded test and as such, could have produced a general conclusion, OP9 concluded his first test of bottle size with a very specific statement: "It did equal out." He seemed to arrive at this conclusion in a hit or miss fashion, talking through it aloud, rather than giving a careful consideration of the possibilities and implications of what he had seen.]

Session Two was the intervention session in which the examiner played a more active role in probing the subject's understanding and leading his thinking. The examiner explained that the chemicals for that session were X, M, and J. She then proceeded with the demonstration and explanation as in the first session. The subject was not, however, questioned about his ideas concerning brightness. As in the first session, the subject listened impassively, without establishing eye contact and without speaking. Following the instructions, the subject proceeded with his tests.

After scanning the bottles with his hand hovering briefly over them, OP9 selected M [illegible] a large bottle with short electrodes and JX in a large bottle with long electrodes for his first test. In response to the examiner's question about selection, the subject replied that he picked M to see if it would light and XJ to see the effect of X on J. [Here the subject seemed to be testing two hypotheses within the same test. The fact that it resulted in a confounded test did not necessarily appear to be significant in view of the subject's explanation for selection.] When asked for a prediction, he replied, pointing to the bottles, that X would "take some out of there" and that he didn't know about M. When the examiner asked what he thought after observing the bulbs lighting, the subject looked down at the bottles, pointing to them as he gave his explanation. He began by saying, "M doesn't have anything in it." When the examiner gave a noncommital, "Um-huh," the subject continued his explanation, halting several times as if he were trying to formulate his answer. When he remarked, "X doesn't take, doesn't look like it took any out of the J," the examiner paused, and questioned the subject. "I don't understand, with what are you comparing it?" The subject replied that he was "comparing the light" and, holding up a bottle from the array, said he could compare it with the test bottle. [Here the subject seemed to be comparing the relative brightness of one of the test bottles with others not necessarily in the stand at the time. In this case, the significant factor was not whether or not the test was confounded, but the subject's use of his memory of the results from previous tests.] The examiner, checking her interpretation of the utterance, looked at the subject and half-questioned, half-stated, "So that was the way, you were remembering from what we had in there before." [Since this was the first test of the session, the examiner and subject seemed to be referring to the subject's memory of the results of the demonstration test.] As he scanned the bottles apparently in search of his next test bottles, the subject indicated that the examiner's statement was accurate with a mumbled, "Yeah."

Following his fourth test, OP9 removed both bottles from the box, set one of them down, and then hesitated as he looked at the remaining bottle in his hand. He started to put the bottle back in the box, but then stopped and placed it on the table. Glancing at the other bottles, he chose a small bottle of J with short electrodes and placed it in the stand. He then returned to the bottle he had just placed on the table, a large bottle of J with long electrodes and placed it in the stand. Although he appeared to be concentrating heavily, his face showed no expression. When asked why he had picked those bottles, OP9 made several starts and stops. [He seemed to be having a great deal of difficulty expressing his thoughts.] At one point, he stopped and grinned slightly while bringing his hand up to his head as if it would help formulate his thought, and mumbled, "Oh, wait, what was I gonna say?" He then stated that he wanted to see how much chemical could get through the shorter stem to charge it up, and that he wanted to compare the lighting of the two chemicals. [Note that in this test he explicitly stated his intention to compare the lighting of the two bottles. In previous tests he often seemed to be drawing information from both bottles separately and applying this to his previous knowledge rather than making comparisons with and drawing conclusions from the bottles used specifically in that test.]

After pressing the switch to light the bulbs, the subject stated that they lit about the same. The examiner began to make a comment, then stopped and asked to see the bulbs light again. Both the examiner and subject leaned forward to

look closely at the bulbs. When the examiner pointed to a bottle and remarked that some people thought that it was a little brighter, the subject looked at the bulbs from several angles before agreeing that it was "some brighter." In response to a question as to why this might be so, he replied that the bottle was brighter because "the stem goes down farther into it, and it can pick up more particles." He also remarked that "J must be really strong," as even the shorter-stemmed bottle was quite bright. The examiner pointed out that there was another difference that might account for the brightness. After a slight clarification of the question, in which the examiner noted the difference in electrode length already mentioned by him, the subject remarked, "Oh, well, the amount that's in this one." When questioned about the effect of bottle size, the subject hesitated, bit his thumb and then stated that it "shouldn't really matter." He went on to explain: "They're both the same chemical and all they need to do is just touch, well, let the electrical current pass through there to light the bulb." When asked how he could prove this, he looked at the bottles and picked up a large bottle of X. He then indicated that bottle size would matter for less powerful chemical, however, because "it needs more to get through."

Following the ninth test, the examiner returned to the demonstration test presented at the beginning of the session. The subject indicated with a mumbled, "yeah" that he remembered what happened, but did not elaborate. The examiner recalled which bottle was brighter and asked OP9 to tell all the things that might be going on that made it so. During his summary, OP9 mentioned that M does not light, J "is pure," and that X is half of J. [He didn't explain or elaborate in these comments, nor did he discuss the effect that each had when in combination with the others. Although he mentioned that one bottle had longer electrodes, he did not explain what affect this might have.]

For tests ten and eleven, the examiner selected bottles to make an unconfounded test of bottle size. In both cases, OP9 predicted that the smaller bottle would light more because the chemicals "are closer together" and "can get to it more." In test ten he perceived his prediction as having been accurate. [It was difficult to know whether or not this actually occurred. In any case, the examiner did not contradict him.]

In test eleven, he again predicted that the smaller bottle would light more. Initially he was sitting back in his chair as he pressed the button, but after a first glance at the bulbs, he leaned forward and moved around to observe the lights and chemicals. Looking less sure of himself, he then released the switch and bent down to observe the bottles. Sitting back again, he put his hand on his chin and smiled as he said, "I don't know why that did that." The examiner also smiled and asked what it looked like to him. He replied that the large bottle lit more than the smaller one. The examiner pressed the switch and repeated his statement, thus offering him the chance to refute it. The subject made no comment in this respect, but hypothesized that perhaps the bottle lit more because it had more in it. He sat back and again put his hand on his chin. In response to the examiner's query, he admitted that he was confused by the results of the test. The examiner explained that perhaps those particular bottles weren't calibrated well enough. She then stated that the same rule would apply for test eleven as for test ten (that smaller bottles light more). The subject, however, responded with the hypothesis that the

larger bottle lit more because the chemicals were better able to move "through" the electrodes. He did not indicate a desire to test this further.

The Third Session was again an observation session. The examiner's interaction with the subject was basically limited to the standard prompts. The session began as the previous sessions, with an instruction/demonstration test by the examiner. When given some paper for notes, the subject declined it with a smile as in the previous session. OP9 took several seconds to scan the bottles, turning some of them as he did so [in order to get a better look at them]. After several seconds, he picked two large bottles, one with F and long electrodes, and one with L and short electrodes. He glanced at the labels once more before placing the bottles into the box. After attaching the wires, he sat back as if waiting for the standard question about his selection. In response to this question, he stated he wanted to see "if they'd light up." When asked what might happen, he made no predictions but grinned and shook his head saying he did not know. Upon observing that only one bulb lit up, he concluded that "F has the chemical that helps light it up," and that he didn't really know about L because it might not have the chemical needed to make the bulb light, or it might not light because of the small electrodes. [This test was significant in that OP9 seemed to have realized that he had made a confounded test and adjusted his conclusion accordingly.]

For the third test, OP9 again scanned the bottles carefully before making a decision on F and D in big bottles with long electrodes. He remarked that he wanted to see which "had more of the chemical in it and which one was more powerful." He predicted that D would light more by pointing to it and saying, "it's more here than this." [It was not clear what the subject was referring to here. As both bottles had the same amount of fluid, it seemed that he might have been referring to the relative chemical strength of D.] He concluded that his prediction was correct, as F "does have extra things in it that take away from the pure thing."

For his fifth test, OP9 selected two large bottles of DF, one with short and one with long electrodes. His stated purpose was to see "which one would light up more" and if the chemical would work as well with the short electrodes. He predicted that the bottle with longer electrodes will "go brighter." Upon confirming this, he stated that "since it (the electrode) goes down farther, the charge can be more, so it lights it up more." [Although this had been previously tested and implied, it was the first time he articulated a general rule governing electrode length.]

Following the sixth test, OP9 commented, "Well, I think I know what I think enough." The examiner suggested that they return to the original bottles. As she leaned over to look for the bottles, the subject selected the correct bottles from the array and helped attach them in the box. When asked about the things involved in making one of the lights brighter than the other, OP9 correctly described the chemicals in terms of their relative strengths. He described F, for example, as "half and half," or half as bright as D. When asked if he thought of some of the other things that might matter for brightness, he replied, "Yeah," and gave the following explanation, pointing to the appropriate bottles as he did so:

> "Stem, it's like an electrode . . . how far down it goes, and F . . needs to go down farther, or it won't light up because it takes a

lot to charge, to make the charge to light the bulb up. D, if there's a longer stem it will light up more. But if there's a short one it'll still light up. And when you put D in, D and L together and F and L, the D and L lights up and the F and L doesn't because it doesn't have enough to light up."

[He clearly understood the relationship between electrode length and the various chemicals as well as the effects of these chemicals on each other.] The examiner then reminded him about one other variable he hadn't yet mentioned, the bottle size, and asked him to tell whether it made a difference. In his response, OP9 indicated that, all other things being equal, a larger bottle would light more than a smaller one. The examiner followed up on OP9's statement by asking him to test this and "tell for sure." After several false starts, OP9 selected the chemical FL with long electrodes in a large bottle and a small one. He again predicted that the larger would light more because "it has a bigger areas so F can get through into more places." After noting that both bottles light the same, he went on to say that because both bottles had the same chemicals, they lit the same. He remarked, "I thought it (bottle size) would make a difference, but it didn't." [It was interesting to note that he had made no previous tests of bottle size in this session. Once it was a confounding factor in a test of chemical, but OP9 did not mention this and in fact appeared not to notice. The subject seemed to be basing his assumption that larger bottles light more on the results of an unconfounded test of bottle size at the conclusion of session two. Although the examiner hinted that the results of that test were not accurate, the subject did not seem to think it necessary to retest this.]

In Summary, several things are important to note about OP9's performance in the three conductivity sessions. First of all, as might be expected, his conclusions and session summaries became more articulate and more complete as the sessions progressed. At the end of the third and final session, for example, his conclusions were much more complex and explicit than in the previous sessions. By the end of the last session he was able to relate the effects of the chemicals in combination with the brightness of the bulbs as well as describe the effects of the various chemicals in combination with the different electrode lengths. In addition, OP9 became more confident and at ease with the situation over time.

Keeping in mind these developments in OP9's explanations and conclusions, an analysis of his general task approach in terms of the relative numbers of confounded and unconfounded tests proves quite interesting. On the basis of his increasingly sophisticated conclusions and explanations, one might expect a similar increase in OP9's preference for unconfounded tests over confounded ones. Instead, just the opposite happened. In the first conductivity session, little more than one-fourth of the tests spontaneously performed by the subject were confounded tests. In the second session, the number of confounded tests rose to slightly more than half, and in the final session, fully two-thirds of the tests performed were confounded. OP9's behavior throughout the conductivity sessions seemed to indicate that he did understand the importance of unconfounded tests for testing hypotheses and for using test results to draw conclusions, even though he did not always do so. In response to examiner prompts for specific tests of a given variable, OP9 used only unconfounded tests, even in the very first conductivity session. OP9 also seemed to understand the implications of confounded tests. Several times

throughout the sessions his statements indicated this when he concluded that a certain effect was due to one of two things. Thus, OP9 may understand the need for unconfounded tests, but didn't always perform them spontaneously. A closer examination of these tests reveals several possible reasons for the increase in the number of confounded tests.

To begin with, OP9 seemed to construct confounded tests not to make comparisons between the two bottles, but to look at each in relation to previously tested bottles. At one point, when explicitly asked whether he did this by the examiner, OP9 confirmed his conscious use of this strategy. By using such a strategy, a subject could theoretically reduce a given number of tests a good deal, provided that he or she could make mental comparisons and keep track of the results. The number of tests performed by OP9 did in fact decrease each session. It is possible that his use of this strategy at least partially accounts for the decline in the number of tests performed per session.

In contrast with OP4, OP9's behavior indicated that he realized that learning is indeed cumulative. He obviously used the information to which he was exposed to guide his behavior. Often, rather than retest a variable, OP9 tended to rely on previous findings and hunches. The most obvious example of this were his tests of bottle size. When faced with information contradictory to his hypotheses or memories, OP9 showed a tendency to distort or ignore such information. He seemed reluctant to make any adjustments in his "memory" strategy and did not test such discrepancies further unless specifically asked to do so by the examiner. Although this strategy proved helpful in some cases, it sometimes got OP9 into trouble, when, as in the bottle size example, he relied too heavily on previous findings. Thus, OP9's reliance on previous findings, his disinclination for further testing to confirm these findings, and his strategy of combining tests all seemed to influence the number and type of tests performed in each session. In terms of his increasingly sophisticated explanations concerning the factors influencing brightness, OP9's improvements throughout the sessions seemed to occur more as a result of his cumulative knowledge about the general task setup than as a result of an increased knowledge and use of unconfounded tests.

The intervention session did not seem to have a great deal of effect on the task approach and test behavior of this subject. The requests for clarification by the examiner, however, proved quite helpful in terms of understanding his thinking. When asked to do so, OP9 provided much more explicit conclusions and explanations. These elaborations by the subject are quite important because they provided information about his level of understanding that would not otherwise have been available. OP9's reluctance to talk, especially in the first session, made such information difficult to obtain.

In general, OP9's comparisons within and outside a given test and his resulting conclusions seem to indicate that he was able to use these comparisons for a better understanding of the influential factors for brightness in the conductivity task. When faced with a contradiction or conflicting information, however, OP9 remained quite rigid in his task approach. He seemed unable to change his thinking and task approach, even when his results would seem to make such an adjustment necessary.

Summary and Preliminary Suggestions for Intervention

Individual Differences in Learning in an IV Task Setting

The three case studies illustrate, in detail, the vast differences in problem-solving and reasoning exhibited by these LD subjects. Despite the fact that their mastery of the IV strategy was similar, on some indices, in both the initial and final rods sessions, their general task approach was quite different. The differences and similarities are summarized below.

OP4 seemed more comfortable and looked more competent in the bending rods task than in conductivity. However, his level of motivation seemed to vary from session to session or moment to moment. He frequently "forgot" to arrive for his sessions and had to be called from his study hall. At times he appeared confused, bored, discouraged, interested, etc. When the examiner became more encouraging and interactive, OP4 seemed to respond with more confidence and enthusiasm. On a number of indices, OP4 showed a steady increase in performance over time in each task setting. Therefore, task familiarity as well as content seemed to have a positive effect on his behavior.

OP4's performance difficulties could be attributed to both verbal and nonverbal factors. His referring expressions were often vague and inconsistent. For example, he used all of the following expressions to refer to the electrodes: iron things, stems, big rods, longer things, metal wire. his descriptions of the outcomes of his tests were also vague, e.g., "J is good and X is not good." He seemed to have difficulty summarizing the results of his tests for the examiner. Because he often omitted several important factors that he had tested, his summaries showed a lack of ability to remember, integrate, or express essential information. Written records of his experiments might have proved helpful, but OP4 decided against keeping such records. In addition, his scores on articulation and use of explanatory principles declined over time.

His nonverbal problems were most apparent in his inappropriate rod replacement and selection strategy and in his initial tendency to choose single-bottle tests in the conductivity task. These poor strategies showed that he has difficulty organizing his activities in unfamiliar task settings. The fact that these inadequate strategies disappeared and then reappeared later in the session suggests a lack of metacognitive awareness about their inadequacy and/or an inflexible task approach.

In light of the many difficulties shown by OP4 prior to the final rods session, his performance in this session was impressive. All evidence of hesitancy, confusion, or uncertainty were gone. His rod selection strategy was thoughtful and systematic, and his verbal expression was explicit. Although gradual mastery of the IV strategy and gradual increases in prediction, explicitness, and intentionality were apparent over the entire intervention period, his behavior in this final session looked qualitatively different from that of any previous session. It is not clear why this dramatic improvement did occur. Some plausible contributing factors may be OP4's greater ease with the rods task, his familiarity with IV task demands in general, and his increased confidence due to the examiner's encouragement.

OP5's approach to both tasks was radically different from that of OP4 in many

respects. Where OP4 was hesitant, OP5 was overly confident. Where OP4 gave vague, brief responses to the examiner's questions, OP5 provided a continuous narration of his activity. Where OP4 used quite concrete referring expressions, OP5 preferred highly abstract, multisyllabic constructs. Unlike OP4, OP5's performance on the conductivity task was often superior to his rods performance. OP5 seemed to enjoy the opportunity the conductivity task provided for the expression and elaboration of hypothetical explanatory constructs.

OP5 seemed to enjoy the attention he received in the intervention sessions, at least initially. He seemed less interested in listening to the examiner than in having the examiner listen to him. By the final conductivity sessions, OP5 became boastful. Several times he implied that the activity would be more appropriate for young children than for someone as scientifically sophisticated as he.

In fact, OP5 showed evidence of early mastery of the IV strategy on the more structured aspects of the rods task: test choice, test evaluation and prompted test construction. Over time, this mastery increased and generalized to the less structured parts of both conductivity and rods tasks. His interest in exploring issues such as explanatory principles that govern conductivity and compensation tests tended to deflate his unconfounded testing scores on the spontaneous portions of the tasks.

OP5, like OP4, showed evidence of both verbal and nonverbal difficulties in these task environments. His nonverbal problems tended to be reflected in his misperceptions of attributes and outcomes. He also showed a tendency to focus on unusual or implausible factors and to ignore more plausible ones. For example, he investigated, repeatedly, the effect of the holes in the stand on differential bending during the first session. This nonstandard variable was rarely tested by other subjects.

OP5's highly verbal task approach increased his scores on a number of task dimensions: general and articulated conclusions; use of explanatory principles. His referring expressions were quite abstract. However, these verbal strengths coincided with some obvious verbal weaknesses. OP5 had a great deal of trouble expressing his ideas in clear, simple, concrete language. His answers were frequently qualified, making it hard to know what he really meant. In addition, he frequently revised his statements in midsentence so he was hard to follow. He seemed to confuse himself as much as he did others. When he took the time to write down some of his conclusions, his explanations and summaries were clearer. Thus, his difficulties with oral expression might be reduced if he were encouraged to refer to written notes.

Finally, OP5 had trouble connecting the experiments he was conducting to the overly general and abstract conclusions he was making. He tended to ignore or distort data that disproved his predictions or conclusions. He was resistant to the examiner's attempts to challenge his thinking. He became distracted by his attempts to develop elaborate, idiosyncratic explanations for differential conductivity.

OP5's eventual mastery of the IV strategy in the final rods session was not surprising given the strengths he displayed in his earlier sessions. However, his mastery of the strategy did not improve his difficulties expressing

himself succinctly and coherently or his tendency to misperceive, distort or ignore, conflicting data. These deficits might need to be the focus of direct instruction in order to see more improvement than could be achieved by the indirect Piagetian clinical method employed here.

OP9 showed many signs that he preferred unconfounded tests early in the intervention even though his proportion of unconfounded spontaneous tests was relatively low. His performance on more structured parts of the task (test choice and test evaluation) was quite good. A large proportion of his spontaneous tests were intentionally confounded in order to investigate how the variables interact. In addition, in the conductivity sessions, his prompted tests were more unconfounded than his spontaneous tests. He also explicitly mentioned the difficulty one has drawing conclusions about confounded tests. Thus, his marked improvement in unconfounded testing between the first and final session seemed to be due to a gradual mastery of the IV strategy over time.

His verbal reasoning was more variable over time. This probably reflected his interest in fully exploring a number of aspects of these tasks. His frequent use of explanatory principles seemed to show that he was trying to develop some reasonable ideas about conductivity from his tests. His terminology also reflected this interest in scientific explanation: acid, additives, charge, particles. His frequent compensation tests showed an interest in the relationship between variables. With time, his general task approach became more confident and his conclusions and summaries more articulate. He also showed the ability to demonstrate that his learning was cumulative by making cross-test generalizations.

Overall, OP9 showed fewer difficulties with these tasks and more sustained interest in them than did either of the other intervention subjects. His major difficulty seemed to be his reliance upon his memory instead of written notes. Thus, he tended not to recheck findings that he thought he had accurately remembered. Also, he seemed to prefer to compare results across tests instead of within a test. (That is, he did not compare the two rods in the stand at the same time, but, instead, compared each to previous rods tested.) This strategy had an obvious flaw, inaccuracy of recall, for which OP9 showed no concern. Another difficulty that he displayed was the tendency to distort or ignore disconfirming evidence.

Therefore, OP9 seemed to make good use of the opportunity to explore the relevant factors in these tasks, to hypothesize scientific constructs to explain the phenomena, and to investigate the interactions between variables. His strong performance could have been improved if he had been encouraged to make use of written notes.

Preliminary Implications for Instruction

The goal of the intervention phase was to explore the utility of a single-subject design and Piagetian clinical interview strategy for the remediation of reasoning and problem solving deficits in LD adolescents. The three case studies detailed and summarized above show that LD adolescents do make different kinds of progress in IV task settings. For some adolescents, their major hurdle may be familiarizing themselves with the task demands, using appropriate terminology, and abandoning inadequate strategies for more

adequate ones. For others, learning to express one's thinking clearly and linking one's ideas to concrete experience may be important. Thus, documenting individual progress over time may be a useful way of identifying these individual differences. However, at this point, it is impossible to assess the generality of the problems and strengths that we saw in these three subjects. In addition, we do not know whether the difficulties they experienced in the two IV tasks are similar to problems they have in other academic areas. Thus, the generalizability of our findings to other subjects and other tasks is unknown at this time.

The single-subject design that we used did provide us with comparable data across time, across tasks and across subjects. Unfortunately, the interpretation of our findings is limited by the lack of comparable normative data. The existing research used a similar design and IV task setting but with much younger normal achieving subjects (4th graders). It suggests that the mastery of the IV strategy is a much more laborious and discontinuous process than we saw with these three LD adolescents. Clearly, more data are needed from older normals and younger LD subjects.

One distinct disadvantage of the single-subject design was the number of sessions devoted to collecting baseline data. The baseline sessions seemed to provide subjects with ample opportunity to master some aspects of the tasks (e.g., the IV strategy). However, when other aspects of task performance were examined, the limitations of our indirect and limited instructional procedure were apparent. Both OP4 and OP5 showed evidence that they could profit from some more direct instruction, not in the IV strategy itself, but in related verbal and noverbal behaviors. For example, OP4 seemed hampered by his vague and inconsistent terminology and OP5 by his incoherent, rambling discourse. These subjects could have benefited from some more focused instruction in vocabulary or in the monitoring of oral expression. In addition, OP4 might have needed some help in organizaing and monitoring his nonverbal problem-solving strategies and OP5 could have used additional instruction in integrating his verbal and nonverbal activities. All three subjects might have benefited from instruction in the use of written records.

GENERAL DISCUSSION AND CONCLUSIONS

The primary purpose of this project was a detailed exploration of the reasoning and problem solving skills of selected subgroups of learning disabled (LD) adolescents. To achieve this purpose, four interrelated issues were addressed: (a) Do LD adolescents evidence difficulty with reasoning and problem solving skills? (b) Do the difficulties vary as a function of the type of learning disability? (c) Can the difficulties seen be distinguished from those evidenced by younger or lower ability controls? (d) Do subjects with difficulties benefit from short-term interventions? These and other related questions were pursued by detailed coding of the verbal and nonverbal behaviors of subjects in the context of two isolation of variables (IV) tasks, bending rods and chemical conductivity. These tasks require the exploration of complex causal events and have been shown to be sensitive to important cognitive changes during the early adolescent years (see the Introduction and Literature Review). Thus, it was assumed that the findings from this study would have important implications for our appreciation of the problems and needs of the LD adolescent population in the area of thinking skills.

Differential Performance of the Three LD Subgroups

Perhaps the most important finding from the project concerns the differential performance of the three LD subgroups on the bending rods task. Although several of the statistical tests did not reveal reliable differences, the differential pattern of performance of the three subgroups was evident across a number of comparisons. On measures of the tendency to construct unconfounded tests, an index of problem solving sophistication used in numerous studies with normal-achieving preadolescents and adolescents, the two Verbal-Performance discrepancy groups (LV and LP) evidenced a level of sophistication above that of the low discrepancy (ND) group and equivalent to that of normal-achieving peers matched for age and fullscale IQ. In contrast, the ND group scored no differently than a group of normal-achieving fourth graders matched for fullscale IQ. This pattern of differential preference for unconfounded tests was evident across both the first and second administrations of the bending rods task.

There was significant improvement in unconfounded testing across the two administrations of the rods task, as has been the case in previous studies of normal adolescents (e.g., Day & Stone, 1982), with little evidence of differential improvement among the three LD subgroups. In all three cases, the improvement tended to be greater than that evidenced by the fourth graders, but less than that evidenced by the ninth grade controls. Only in the LP group did the level of improvement approach that of the ninth graders. On the basis of recent work on the use of metacognitive strategies in LD children (e.g., Torgesen & Licht, 1983), one might have expected the LD subgroups to show more improvement across administrations than was shown by normal-achieving subjects. Such a pattern has been widely interpreted as indicating that LD children may have access to higher-level cognitive strategies but fail to make use of them spontaneously. This pattern has been found on measures of strategic memory behaviors by several researchers (cf. Torgesen & Licht, 1983) and on measures of inferential comprehension (Wong, 1979). However, such an explanation for the failure of some subjects to make consistent use of the IV strategy is not supported by the present data. Subjects in the three LD subgroups tended to make more progress than fourth graders but similar, or slightly less progress than their same-age peers.

Although the test choice and test evaluation questions are additional indices of the preference for using unconfounded tests, these measures yielded a different pattern of differential group performance. This different pattern was the result of the markedly lower performance of the LV group, which performed more like the ND than the LP group on these measures. Both the test choice and test evaluation measures require subjects to state a preference for conducting unconfounded tests, but they differ from the test construction measures in that they also require the subjects to justify their preference by referring to the confounding variable in the rejected test pair. The relatively poor performance of the LV subjects on these measures serves to distinguish their mastery of the IV strategy from that of the LP subjects: the LV subjects were likely to construct unconfounded tests but were less likely to argue that such tests are preferable. (The possible role of language difficulties in explaining these results is discussed in a later section.

The results of the analyses of performance on the conductivity task suggest

that the findings with respect to unconfounded testing described above are not artifacts of the specific rods task environment. The pattern of group differences evident during the second test construction phase of the rods task were again evident one week later on the conductivity task. Also the correlations between subjects' scores on the two tasks reveal moderate correspondences. The correlations with the Concept Formation subtest of the Woodcock-Johnson (Woodcock & Johnson, 1977) suggest that the problems seen here bear some relationship to those evident on a widely-used standardized measure, but the moderate size of the correlations suggest that complex problem solving tasks such as the rods and conductivity tasks may highlight behaviors not tapped by existing standardized tests.

As anticipated, the discriminant analyses of the data derived from the reasoning and examiner guidance codes provide valuable information concerning subgroup differences to complement that obtained from the analyses of unconfounded testing. In general, the results of comparisons of the three LD groups highlight differences in reasoning sophistication between the LV and LP groups. During the initial test construction phase, the LV subjects differed from their LP peers in the quality of the predictions they made concerning these tests. The LV subjects tended to evidence less intentionality and to make less explicit predictions. On these variables, the ND subjects scored between the two high discrepancy groups. During the second test construction phase, the LP group again showed stronger reasoning skills than the LV group, with the ND group again scoring between the two. During TC2, the LP subjects were more explicit in their connections between variables and outcomes (high articulation of conclusion) and made more use of explanatory principles to account for their findings.

Thus, the data from the entire set of analyses provide the following picture with respect to the differences among the three LD groups. The LV and LP groups show an equal and greater tendency to conduct unconfounded tests than the ND group when exploring the factors influencing relative bending. However, the LP group was better able to justify their unconfounded tests (test choice and evaluation) and described the nature and outcome of their tests more clearly than the LV subjects, with the ND subjects scoring between these two extremes.

Comparisons to Normal Controls

Inclusion of the normal comparison groups allowed a determination of the extent to which the performance of the LD subgroups was distinct from that typically found. For example, the performance of the LP group is consistently strong on all measures relative to their other LD peers, but how do they compare to the normal-achieving age-mates matched for IQ? Here again, the findings for this LD subgroup are remarkably consistent. On all measures of unconfounded testing, the LP group scored equivalently to their same-age controls and significantly above the level of the fourth graders. Similarly, the discriminant analyses indicate that the LP subjects are more like ninth than fourth grade controls. During TC1, the LP and N9 subjects score high relative to fourth graders on measures of the generality of conclusions and the use of explanatory principles. During TC2, the two groups out-perform the fourth graders on measures of explicitness of conclusions and compliance with examiner's requests. Finally, the classification tables generated by the

discriminant analyses reveal that the LP subjects are often misclassified as normal ninth graders.

If the performance of the LV and ND subjects is poor in certain respects, can they be distinguished from younger normal subjects? While the ND group shows a level of unconfounded testing equal to that of fourth graders on all measures, their task approach can be distinguished from that of their younger peers on certain reasoning measures. During TC1, the ND subjects show a level of use of general conclusions and explanatory constructs more like that of ninth graders than that of fourth graders. However, during TC2, their task approach is again more like that of the fourth graders than that of their ninth grade peers. The discriminant analysis reveals a level of inexplicit conclusions and poor compliance with examiner requests equal to that of the fourth graders. This seeming inconsistency is due to the fact that the ninth graders begin to adopt a more detailed approach to examining the causal factors during TC2, while the ND group maintains a highly general, nonspecific orientation (see the task approach Summary).

The findings for the LV group are also mixed. On measures of unconfounded test construction, the LV subjects perform more like ninth graders than fourth graders, but on measures of test justification, they perform like fourth graders. Similarly, their reasoning performance shows some similarities with both age groups; however, overall, their performance is more like that of the fourth graders. During TC1, the LV subjects draw general conclusions at the same frequency as ninth graders, but they resemble fourth graders in their infrequent use of explanatory principles and noncompliance with examiner requests. Their expression of intentionality in describing their tests was lower than that of the fourth graders. In TC2, the LV subjects draw conclusions which are more general than those of the ninth grade controls, but their conclusions are less explicit, and they are again less likely to comply with examiner requests. In the classification analyses for both TC1 and TC2, the LV subjects are often misclassified as fourth graders.

The inclusion of the low IQ normal ninth grade group was intended to address the question of whether or not the LD subgroups were merely performing similarly to normal-achieving subjects of lower overall ability. Unfortunately, the difficulties encountered in recruiting these subjects (see Subjects) made adequate comparisons impossible. It was not possible to match their mean fullscale IQ to that of the lower subscale of the two high discrepancy groups, as originally planned. Furthermore, the small sample size (n=6) made statistical comparisons unwise. The analyses of unconfounded testing scores did suggest, however, that this issue should be pursued in the future. The performance of the ND subgroup on the measures of preference for unconfounded tests was no different from that of the low IQ group. This did not appear to be the case for the other two subgroups, whose performance was between that of the two IQ comparison groups. Firm conclusions regarding this issue should await additional data from larger samples and from discriminant analyses of task approach, however.

The Nature of Reasoning and Problem Solving Difficulties

How should we interpret the present findings? When designing this project, we drew on a pilot study conducted in a clinic setting by one of the authors (Stone, 1981) to aid in defining the LD subgroups and in generating

predictions. The data from that study led us to expect that the two high discrepancy groups would perform poorly on the rods task, while the ND group would perform much like same-age controls. Clearly, these expectations were not borne out. While there are many possible explanations, including minor differences in task procedures and coding criteria, the most reasonable explanation involves the specific samples used. There are two major differences between the samples in the two studies. First, the Stone (1981) study was based on a university clinic sample, while the present study used a public school sample. Second, the subgroups in the pilot study were defined clinically, while those in the present study were defined using numerical criteria.

Fortunately, data collected from the teacher questionnaires allow a comparison of the major difficulties evidenced by subjects in the present subgroups. As discussed earlier in this report (see Subjects), the results of an analysis of the teachers' reports revealed that the subjects in the ND subgroup manifest problems with thinking skills and language comprehension. Thus, their difficulties on the rods task are actually consistent with teacher reports of their primary difficulties. In contrast, the subjects in the pilot study who had primarily decoding problems were largely free of difficulties with thinking skills. The other two LD subgroups in the present sample bear more resemblance to those discussed by Stone (1981). Subjects in the LV group tended to have language and reading problems, while those in the LP groups had exclusively nonverbal problems.

The relatively strong preference for constructing unconfounded tests on the part of the LV subjects is in marked contrast to their performance on the other measures. However, the reasoning measures are clearly language dependent. Also, it is noteworthy that the LV subjects consistently showed poor performance on the measure of compliance with examiner requests. One possible interpretation of this noncompliance could be that the LV subjects failed to comprehend the nature of the examiner's requests. Thus, many of the problems evidenced by the LV subjects may relate directly to their difficulties with receptive and expressive language. Even the highly structured test choice and evaluation questions required comprehension of a complexly-phrased question and the ability to express one's thoughts easily. On the basis of past research, however, one might have expected the LV subjects to evidence difficulties with unconfounded testing as well. Recent studies by Friedman (1984), Johnston and Ramstead (1983), and Kahmi (1981) have provided evidence of conceptual disorders in younger language-disabled children. Such disorders might reasonably be expected to interfere with the construction of the IV strategy. In this context, it is possible to interpret the poor performance of the LV group on the test choice and evaluation questions as evidence of a failure to understand fully the need to control variables. Subjects' performance on these questions was broken down into the judgment they provided (the test they chose as "better" for the test choice questions, or their yes/no response for the test evaluation questions) versus the justification they gave for that judgment. Separate analyses of these scores yielded identical patterns of poor performance. Thus, the LV subjects' low scores on the test choice and test evaluation questions cannot be attributed to difficulties in verbal expression. Also, informal analyses of their justifications suggest that they understood the examiner's questions. Existing data from normal-achieving children (Day & Stone, 1982) suggest that a discrepancy between the construction of unconfounded tests and the

conviction that they are necessary is characteristic of developmentally younger children. Thus, the LV subjects appear to have a less sophisticated understanding of the need to control variables than their test construction scores would suggest, and their poor performance does not appear to be directly attributable to their language problems. A similar analysis of the direct role of language problems in the poor performance of the LV subjects on the reasoning and examiner compliance measures is beyond the scope of this report. However, the data from the test choice and test evaluation questions suggest that this poor performance may in fact reflect genuine difficulties in thinking as well.

The performance of the LP group is the most puzzling of the three LD groups. Unfortunately, this puzzle is not clarified by teacher reports of the difficulties evidenced by these subjects. Several subjects in this group were reported to have difficulties with thinking skills, yet the performance of this group was consistently the strongest of the three LD groups. On the basis of past research, one might have expected to see strong performance on the more verbal reasoning measures, as was the case, but the strong performance of the LP group on the measures of unconfounded testing is somewhat surprising. In one of the very few studies of thinking skills in LD children with nonverbal disorders, Schmid-Kitsikis (1972) found that such children could deal quite effectively with Piagetian concrete operational tasks (e.g., seriation) if they were allowed to explain what should be done, but that these same subjects could not execute their own suggestions. By analogy, one might expect subjects in the LP group to explain the need to control variables but fail to do so effectively. This was not the case in the present sample.

In the intervention phase, we found that LD adolescents who show initial difficulties in reasoning and problem-solving on a complex task can make substantial improvement over a relatively brief period of time. This improvement can also be achieved in the absence of explicit instruction. However, the generality of these findings to other LD subjects and to other tasks remains to be established in future research. The verbal and nonverbal difficulties that persisted in some of the subjects throughout the intervention may require a more direct instructional approach. Future research in which indirect, Piagetian clinical interview techniques are supplemented with more explicit training may help identify the most effective instructional strategies for LD adolescents. Comparable intervention data from normal achieving average ability and low ability subjects and younger normal subjects would also provide information necessary for designing appropriate instructional materials. Given the variability that we found in our three intervention subjects, future research on the remediation of thinking skills in LD adolescents should continue to employ a single-subject design so that individually-tailored instruction can be developed.

Taken as a whole, the data summarized above provide evidence of difficulties with reasoning and problem solving skills in many LD adolescents. However, they also provide a clear indication that not all LD adolescents have such difficulties, and that the specific nature of the difficulties varies as a function of the type of learning disability and of the task demands. These findings provide some corroboration for the numerous reports of LD clinicians (see the Introduction) and for the reports of the LD teachers working with the

subjects used in the present study (see Documentation of Need for Research on Thinking Skills and Subjects).

Educational and Theoretical Implications

What implications should we draw from these findings concerning the reasoning and problem solving difficulties of LD adolescents? One clear implication concerns the need for caution in drawing general conclusions about the difficulties seen. While many of the LD subjects evidenced difficulties in reasoning and problem solving in the IV tasks, many did not, and those who did varied in the nature of the problem they evidenced. Thus, global assertions about thinking skill deficiencies in adolescents as a whole should be avoided. On the other hand, although the problems seen were not shared by all LD adolescents, they were sufficiently frequent and serious to warrant special attention. The fact that problems with thinking skills were often corroborated by the subjects' resource teachers suggests that the difficulties highlighted in the IV task setting are general ones which extend to academic settings. Diagnosticians need to be alert to these potential problems. Ultimately, a behavior rating scale such as the Teacher Rating Scale described elsewhere in this report may prove useful in allowing diagnosticians and/or LD resource teachers to identify such difficulties.

A related implication concerns the generality of the problems evidenced in the IV tasks and the forms they might take in other settings. As mentioned before, the fact that there was some correspondance between the LD resource teachers' reports of thinking skill difficulties and the relative performance of the three LD subgroups provides some general corroboration of the present findings. The low, but reliable correlation to the Concept Formation subtest of the Woodcock-Johnson is also encouraging. Thus, the difficulties in reasoning and problem solving seen in the present context may represent general problems which interfere in other areas of functioning.

How might such problems manifest themselves? Perhaps the most obvious situations in which to expect similar difficulties would involve relatively unstructured, open-ended problem solving situations. These would occur in science laboratory courses and in some math courses (planning geometry proofs). They would also occur in English or social studies classes, where students have to organize term papers or develop arguments. In these situations, subjects who failed to adopt the systematic data-gathering skills evidenced by a preference for unconfounded testing might appear to be disorganized in approaching the academic activities mentioned. Subjects who evidenced low scores on the reasoning measures (e.g., intentionality, generality, or explanations) might be expected to evidence difficulties in thinking critically about their own ideas, or those of others. They might also fail to take an appropriately abstract view of new concepts or arguments.

While these descriptions may seem to be plausible extrapolations from the problem-solving and teacher-report data analyzed in the present study, it should be remembered that this project has provided no direct evidence that such extrapolations are warranted. The present findings do, however, provide some empirical justification for the potential value of further examination of these issues.

A general question which is raised naturally by the present findings relates to the issue of delay vs. difference in cognitive development (see Literature Review). If certain subgroups of the LD adolescent population show levels of performance equal to those of younger normal children, should one conclude that their cognitive development is merely delayed? Unfortunately, the results summarized above suggest that this question is not a simple one. While the ND and LV subjects perform like the fourth graders on some measures, their performance is more like that of their ninth grade peers on other measures. Furthermore, the specific aspects of performance which cause these subjects difficulty are not the same. The LV subjects resemble ninth graders on the relatively nonverbal test construction measures, but are more like fourth graders on the more verbal measures. The ND subjects resemble the fourth graders on all measures of unconfounded testing but are more like ninth graders on certain reasoning measures.

Clearly, these LD adolescents are not functioning in a general sense like normal-achieving fourth graders. To go beyond this statement, however, would require more information concerning the normal developmental progression between these two age extremes. It would also be necessary to call upon a a general theoretical conception of the behaviors under development. Neither a general stage model (e.g., Inhelder & Piaget, 1958) nor an information processing model (e.g., Case, 1974; Siegler, 1981) is as yet detailed enough to be of help in this case. Additional research with normal subjects between the ages of the fourth and ninth graders, and with LD samples evidencing wide inconsistencies in task performance might serve to illuminate these issues. Both the LV and ND groups showed signs of a discrepancy between the unconfounded testing and verbal reasoning measures, but the relative performance on the two sets of measures was different. Additional information concerning the incidence and exact nature of such discrepancies in both normal and LD samples might serve to refine our understanding of how complex cognitive skills, such as the IV strategy, are constructed out of component verbal and nonverbal skills. In turn, such an understanding could lead to improved techniques for the assessment and remediation of thinking skills in exceptional populations.

Limitations of the Present Study and Need for Future Research

As with any initial study in a new area of research, the findings discussed above are in need of replication and extension before their full value can be realized. In this section we will address a few of the immediate implications for future research.

Probably the most important need is to extend the data base and analyses to additional samples . The fourth and ninth grade normal-achieving samples used in the present study provided useful pictures of two ends of a developmental continuum for initial comparisons to the performance of LD adolescents. However, further conclusions concerning the nature and severity of the problems encountered by the LD samples would require more detailed normative information: Such information would be provided by a sample of sixth graders and by larger samples of fourth and ninth graders. One specific group of normal controls which merits further investigation are lower ability students. Comparisons of the LD groups in the present study to the low IQ control group were hampered by our failure to identify a sufficient number of such subjects, but the analyses did suggest that future comparisons might prove useful.

These data would allow a better characterization of developmental and individual differences in performance in the normal population.

Additional samples of LD students would also provide useful information. Two findings regarding the present samples were unexpected and merit further investigation. First, as discussed earlier, the strong skills displayed by the LP subgroup were unexpected. A second, larger sample of such subjects is clearly needed. Second, the poor performance of the ND subgroup was also unexpected. The confirmation from the LD teacher questionnaires of difficulties with thinking skills in this sample suggests that the findings are valid for these subjects, but their generality is in need of further investigation. Replication of this poor performance would lead to additional questions concerning the cause of these difficulties. Such questions could be pursued via investigation of possible differential qualification criteria for special educational services at the elementary and secondary levels. (Are poor readers without overt language disorders or thinking problems simply no longer served at the secondary level?) Alternatively, these questions will be addressed via a longitudinal study of thinking skills in LD children during the preadolescent years. Related to this latter research direction is the need for the testing of older LD adolescents. Such data would provide a more adequate assessment of the developmental delay vs. difference issue with respect to higher cognitive skills.

Additional needs for future research are related to the measures and analyses used in the present study. While the isolation of variables tasks provide a rich sample of reasoning and problem solving behavior and warrant additional use, procedures used might be modified somewhat, and additional tasks should be used to supplement the data base. There are two major procedural changes which warrant attention. First, although valuable for purposes of standardization of procedures, the limitations placed on the examiners in terms of permissable questions resulted at times in unnecessarily ambiguous data. Allowing the examiner to ask the subjects to clarify their remarks via nondirective questions would make data interpretation easier. Second, the addition of questions pertaining to causal explanations would provide useful information concerning intuitive and formal scientific reasoning.

In additional to the needs for procedural refinements, future studies might include additional measures of conceptual and reasoning skills in order to assess the generality of the findings. Inclusion of additional standardized measures of conceptual, reasoning, and problem solving skills would also serve to examine the possibilities for screening indices. The data from the Woodcock-Johnson Concept Formation subtest indicated a moderate degree of overlap with measures of unconfounded testing. Additional work is necessary, however, to determine whether that and other standardized measures might serve to identify subjects likely to evidence the types of difficulties seen on the IV tasks. Inclusion of a modified version of the Teacher Rating Scale in such a study would serve to compare the relative utility and efficiency of standardized and informal screening devices.

The final implications for future research to be addressed concern the need for additional data analysis procedures to supplement those used here. The present analyses focussed primarily on static assessments of behavior. However, the differences noted between the two test construction phases and the progress noted during the pilot intervention sessions suggest that

analyses focussed more directly on microdevelopmental changes within and across several sessions would be highly informative. Both qualitative descriptive procedures and formal analyses of sequential probabilities might prove useful in highlighting differential progress.

In additional to more sensitive analyses of the performance of the various samples described above, future efforts should include empirical efforts to derive subgroups defined in terms of reasoning and problem solving performance. Multivariate procedures such as cluster analysis have been used successfully to address such issues with respect to subgroups of poor readers and would be directly applicable to data such as those described here if a larger number of subjects were available and if a priori subgroup distinctions were collapsed.

Implementation of the additional research suggestions discussed here would have implications for the assessment and remediation of LD adolescents and for programming issues at the secondary level. Refinement and further validation of the Teacher Rating Scale and additional investigation of the utility of a battery of standardized instruments for screening for the difficulties evidenced in IV tasks would lead to specific suggestions for assessment procedures. The data from additional samples would serve to identify those subgroups in most need of screening and to pinpoint the specific difficulties one might expect to find. The additional normative data from preadolescent and adolescent controls would refine our information concerning the development of reasoning and problem solving skills. Together with the data from additional multiple session designs, this information would lead to more specific suggestions regarding goals and procedures for intervention. Finally, additional information concerning the incidence, severity, and generality of reasoning and problem solving difficulties in the LD adolescent population would have implications for the importance we should attach to such issues in designing special educational programs and in counselling students regarding mainstream course work.

REFERENCES

deAjuriaguerra, J., Jaeggi, A., Guinard, F., Kocher, F., Maquard, M., Roth, S., & Schmid, E. (1965). Evolution et prognostic de la dysphasie chez l'enfant [Development and prognosis of dysphasia in children]. La Psychiatrie de l'Enfant, 8, 391-452.

deAjuriaguerra, J., Jaeggi, A., Guignard, F., Kocher, F., Maquad, M., Roth, S., & Schmid, E. (1976). The development and prognosis of dysphasia in children. In D. M. Morehead & A. E. Morehead (Eds.), Normal and deficient child language (pp. 345-385). Baltimore: University Park Press.

deAjuriaguerra, J., Jaeggi, A., Guinard, F., Kocher, F., Maquard, M., Paunier, A., Quinodoz, D., & Siotis, E. (1965). Organisation psychologique et troubles du developement du language: Etude d'un groupe d'enfants dysphasiques [Psychological organization and problems of language development: A study of a group of dysphasic children]. In Problemes de psycholinguistique, Vol. 6. Paris: Presses Universitaires de France.

Andersson, K. E., Richards, H. C., & Hallahan, D. P. (1980). Piagetian task performance of learning disabled children. Journal of Learning Disabilities, 13, 501-505.

Andrew, J. M. (1974). Delinquency, the Wechsler P. sign and the I-level system. Journal of Clinical Psychology, 30, 331-335.

Arbitman-Smith, R., & Haywood, H. C. (1980). Cognitive education for learning-disabled adolescents. Journal of Abnormal Child Psychology, 8, 51-64.

Ausubel, D. P., & Ausubel, P. (1966). Cognitive development in adolescence. Review of Educational Research, 36, 403-413.

Beck, H. S., & Lam, R. H. (1955). Use of the WISC in predicting organicity. Journal of Clinical Psychology, 11, 154-158.

Belmont, L., & Birch, H. G. (1966). The intellectual profile of retarded readers. Perceptual and Motor Skills, 22, 787-816.

Black, F. W. (1974). WISC verbal performance discrepancies as indicators of neurological dysfunction in pediatric patients. Journal of Clinical Psychology, 30, 165-167.

Blalock, J. (1977). A study of conceptualization and related abilities in learning disabled and normal preschool children. Unpublished doctoral dissertation, Northwestern University, Evanston.

Blasi, A., & Hoeffel, E. C. (1974). Adolescence and formal operations. Human Development, 17, 344-363.

Bredderman, T. A. (1973). The effects of training on the development of the ability to control variables. Journal of Research in Science Teaching, 10, 189-200.

Camp, G. W. (1966). WISC performance of acting-out and delinquent children with and without EEG abnormality. Journal of Consulting Psychology, 30, 350-353.

Capon, N., & Kuhn, D. (1979). Logical reasoning in the supermarket: Adult females' use of a proportional reasoning strategy in an everyday context. Developmental Psychology, 15, 450-452.

Case, R. (1972). Validation of a neo-Piagetian mental capacity construct. Journal of Experimental CHild Psychology, 14, 187-302.

Case, R., & Fry, C. (1973). Evaluation of an attempt to teach scientific inquiry and criticism in a working class high school. Journal of Research in Science Teaching, 10, 135-142.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Copeland, A. P., & Weissbrod, C. S. (1983). Cognitive strategies used by learning disabled children: Does hyperactivity always make things worse? Journal of Learning Disabilities, 16, 473-477.

Danner, F. N., & Day, M. C. (1977). Eliciting formal operations. Child Development, 48, 1600-1606.

Day, M. C. (1978). Adolescent thought: Theory, research, and educational implication. Washington, DC. Report prepared for the National Institute of Education.

Day, M. C., & Stone, C. A. (1982). Developmental and individual differences in the use of the control-of-variable strategy. Journal of Educational Psychology, 74, 749-760.

Dean, R. S. (1977). Patterns of emotional disturbance on the WISC-R. Journal of Clinical Psychyology, 33, 486-490.

Deshler, D. D. (1978). Psychoeducational aspects of learning-disabled adolescents. In L. Mann, L. Goodman, & J. L. Wiederhold (Eds.), Teaching the learning-disabled adolescent (pp. 47-74). Boston: Houghton Mifflin.

Dulit, D. (1972). Adolescent thinking a la Piaget: The formal stage. Journal of Youth and Adolescence, 1, 281-301.

Erwin, J., & Kuhn, D. (1979). Development of children's understanding of the multiple determination underlying human behavior. Developmental Psychology, 15, 352-353.

Feuerstein, R., Rand, Y., Hoffman, M., & Miller, R. (1979). Instrumental enrichment. Baltimore: University Park Press.

Fincham, F. (1979). Conservation and cognitive role-taking ability in learning disabled boys. Journal of Learning Disabilities, 12, 25-31.

Fincham, F. (1982) Piaget's theory and the learning disabled: A critical analysis. In S. Modgil & C. Modgil (Eds.), Jean Piaget: Consensus and controversy (pp. 369-390). New York: Praeger.

Forman, E. A. (1981). The role of collaboration in problem-solving in children. Unpublished doctoral dissertation, Harvard University, Boston.

Frick, T., & Semmel, M.I. (1978). Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 48(1), 157-184.

Friedman, J. (1984). Classification skills in language disordered, deaf, and normal preschoolers: A study in language and conceptual thought. Unpublished doctoral dissertation, Northwestern University.

Goldschmid, M., & Bentler, P. (1968). Concept assessment kit - conservation. San Diego: Educational and Industrial Testing Service.

Havertape, J. F. (1976). The communication function in learning disabled adolescents: A study of verbalized self-instructions. Dissertation Abstracts International, 37, 1489-A.

Havertape, J. F., & Kass, C. E. (1978). Examination of problem-solving in learning disabled adolescents through verbalized self-instruction. Learning Disabilities Quarterly, 1, 94-100.

Henning, J., & Levy, R. (1967). Verbal-performance I.Q. differences of white and negro delinquents on the WISC and WAIS. Journal of Clinical Psychology, 23, 164-168.

Holroyd, J., & Wright, F. (1965). Neurological implications of WISC verbal-performance discrepancies in a psychiatric setting. Journal of Consulting Psychology, 29, 206-212.

Hopkins, K. D. (1964). An empirical analysis of the efficiency of the WISC in the diagnosis of organicity in children of normal intelligence. Journal of Genetic Psychology, 105, 163-172.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. Psychological Bulletin, 95, 156-171.

Huelsman, C. B. (1970). The WISC subtest syndrome for disabled readers. Perceptual and Motor Skills, 30, 535-550.

Hurd, P. D. (1978). Final report of the national science foundation early adolescence panel meeting, April 30, May 1, 2, 3, 1978. In Early adolescence: Perspectives and recommendations. Washington, D.C.: National Science Foundation, Directorate for Science Education.

Inhelder, B., (1966). Cognitive development and its contributions to the diagnosis of some phenomenon of mental deficiency. Merrill-Palmer Quarterly, 12, 299-319.

Inhelder, B. (1968). The diagnosis of reasoning in the mentally retarded. New York: Chandler.

Inhelder, B. (1976). Observations on the operational and figurative aspects of thought in dysphasic children. In D. M. Morehead & A. E. Morehead (Eds.), Normal and deficient child language (pp. 335-343). Baltimore: University Park Press.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.

Inhelder, B., & Siotis, E. (1963). Observations sur les aspects operatifs des enfants dysphasiques. In Problemes de psycholinguistique (Vol. 6, pp. 143-153). Paris: Presses Universitaires de France . Reprinted in translation in Morehead, D. M., & Moorehead, A. B. (Eds.) (1976), Normal and deficient child language (pp. 335-343). Baltimore: University Park Press.

Jackson, S. (1965). The growth of logical thinking in normal and subnormal children. British Journal of Educational Psychology, 35, 255-258.

James, K. (1975). A study of the conceptual structure of measurement of length in LD and normal children. Dissertation Abstracts International, 36, 4401A. (University Microfilms No. 75-29665,228)

Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior Change: Methodology, concepts and practice (pp. 7-67). Champaign, IL: Research Press.

Johnston, J. R. (1982). The language disordered child. In N. J. Lass, L. V. McReynolds, J. L. Northern, & D. Yoder (Eds.), Speech, language, and hearing: Vol. 2. Pathologies of speech and language (pp. 780-797). Philadelphia: W. B. Saunders.

Johnston, J. R., & Ramstad, V. (1983). Cognitive development in preadolescent language-impaired children. British Journal of Disorders of Communication, 18, 49-55.

Kahn, J. (1976). Training EMR and intellectually average adolescents of low and middle SES for formal thought. American Journal of Mental Deficiency, 79, 397-403.

Kamhi, A. G., (1981). Nonlinguistic symbolic and conceptual abilities of language-impaired and normally developing children. Journal of Speech and Hearing Research, 24, 446-453.

Karmiloff-Smith, A., & Inhelder, B. (1975). If you want to get ahead, get a theory. Cognition, 2, 195-212.

Kaufman, A. S. (1976). Verbal-performance IQ discrepancies on the WISC-R. Journal of Consulting and Clinical Psychology, 44, 739-744.

Kaufman, A. S. (1979). Intelligent Testing With the WISC-R. New York: Wiley.

Kaufman, A. S. (1981). The WISC-R and learning disabilities assessment: State of the art. Journal of Learning Disabilities, 14, 520-526.

Keating, D. P. (1980). Adolescent thinking. In J. Adelson (Ed.), Handbook of adolescent psychology (pp. 211-246). New York: John Wiley.

Kirchner, D. M., & Skarakis-Doyle, E. (1983). Developmental language disorders: A theoretical perspective. In T. M. Gallagher & C. A. Prutting (Eds.), Pragmatic assessment and intervention issues in language (pp. 215-235). San Diego: College-Hill Press.

Kirk, R.E. (1968). Experimental design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole.

Klees, M., & Lebrun, A. (1972). Analysis of the figurative and operative processes of thought of 40 dyslexic children. Journal of Learning Disabilities, 5, 389-396.

Knight-Arest, I., & Reid, D. K. (1979, February). Peer interaction as a catalyst for conservation acquisition in normal and learning disabled children. Paper presented at the Ninth Annual Conference on Piagetian Theory and the Helping Professions, Los Angeles.

Kronick, D. (1978). An examination of psychosocial aspects of learning disabled adolescents. Learning Disabilities Quarterly, 1, 86-93.

Kuhn, D. (1979). The significance of Piaget's formal operations stage in education. Journal of Education, 161, 34-50.

Kuhn, D., & Angelev, J. (1976). An experimental study of the development of formal operational thought. Child Development, 47, 697-706

Kuhn, D., & Brannock, J. (1977). Development of the isolation of variables scheme in an experimental and "natural experiment" context. Developmental Psychology, 13, 9-14.

Kuhn, D., & Ho, V. (1977). The development of schemes for recognizing additive and alternative effects in a "natural experiment" context. Developmental Psychology, 13, 515-516.

Kuhn, D., Ho, V., & Adams, C. (1979). Formal reasoning among pre- and late adolescents. Child Development, 50, 1128-1135.

Kuhn, D., & Phelps, E. (1979). A methodology for observing development of a formal reasoning strategy. New Directions for Child Development, Number 5: Intellectual Development Beyond Childhood (pp. 45-58). San Francisco: Jossey-Bass.

Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. In H. Reese & L. Lipsitt (Eds.), Advances in Child Development and Behavior (Vol. 17). New York: Academic Press.

Larson, M. A., & Dittmann, F. E. (1975). Compensatory education and early adolescence: Reviewing our national strategy (Research Report EPRC 2158-7). Menlo Park, CA: Stanford Research Institute, Educational Policy Research Center.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. Journal of Research in Science Teaching, 15, 11-24.

Lawson, A. E., Blake, A. J. D., & Nordland, F. H. (1975). Training effects and generalization of the ability to control variables in high school biology students. Science Education, 59, 387-396.

Lawson, A. E., & Wollman, W. T. (1976). Encouraging the transition from concrete to formal cognitive functioning--An experiment. Journal of Research in Science Teaching, 13, 413-430.

Linn, M. C. (1978a). Final report of NIE-SRCD conference on future research in adolescent reasoning (Contract No. NIE-P-78-0023). Washington, DC: National Institute of Education.

Linn, M. C. (1978b). Influence of cognitive style and training on tasks requiring the separation of variables schema. Child Development, 49, 874-877.

Linn, M. C., & Levine, D. I. (1978). Adolescent reasoning: Influences of question format and type of variables on ability to control variables. Science Education, 62, 377-388.

Lister, C. M. (1970). The development of a concept of volume conservation in ESN children. British Journal of Educational Psychology, 40, 55-64.

Lister, C. M. (1972). The development of ESN's children's understanding of conservation in a range of attribute situations. British Journal of Educational Psychology, 42, 14-22.

Lovell, C. (1961). A follow-up study of Inhelder and Piaget's The Growth of Logical Thinking. British Journal of Psychology, 52, 143-153.

Lovell, K., & Shayer, M. (1978). The impact of the work of Piaget on science curriculum development. In J. M. Gallagher & J. A. Easley (Eds.), Knowledge and development: Vol. 2. Piaget and education (pp. 93-138). New York: Plenum.

Martorano, S. C. (1977). A developmental analysis of performance on Piaget's formal operations tasks. Developmental Psychology, 13, 666-672.

Martorano, S. C., & Zentall, T. R. (1980). Children's knowledge of the separation of variables concept. Journal of Experimental Child Psychology, 30, 513-526.

Matarazzo, J. D. (1972). Wechsler's measurement and appraisal of adult intelligence (5th ed.). Baltimore: Williams and Wilkins.

McFarland, T., & Grant, F. (1982). Contribution of Piagetian theory and research to an understanding of children with learning problems. In. S. Modgil & C. Modgil (Eds.), Jean Piaget: Consensus and controversy (pp. 391-405). New York: Praeger.

McHugh, A. F. (1963). WISC performance in neurotic and conduct disturbances. Journal of Clinical Psychology, 19, 423-424.

Meltzer, L. J. (1978). Abstract reasoning in a specific group of perceptually impaired children: Namely, the learning-disabled. Journal of Genetic Psychology, 132, 185-195.

Neimark, E. D. (1975). Intellecutal development during adolescence. In F. D. Horowitz (Ed.), Review of child development research (Vol. 4 pp. 541-594). Chicago: University of Chicago Press.

Neimark, E. D. (1979). Current status of formal operations research. Human Development, 22, 60-67.

Neimark, E. D. (1980). Intellectual development in the exceptional adolescents as viewed within a Piagetian framework. Exceptional Education Quarterly, 1, 47-56.

Peel, E. A. (1971). The nature of adolescent judgment. New York: John Wiley.

Prentice, N. M., & Kelly, F. J. (1962). Psychological testing in the correctional institution: Another viewpoint. Crime and Delinquency, 10, 263-268.

Pulos, S. M., & Linn, M. C. (in press). Generality of the controlling variables scheme in early adolescence. Journal of Early Adolescence.

Reed, J. C. (1967). Reading achievement as related to differences between WISC verbal and performance IQs. Child Development, 38, 835-840.

Reid, D. K. (1978). Genevan theory and the education of exceptional children. In J. Gallagher and J. Early, Jr. (Eds.) Knowledge and Development (Vol. 2, pp. 199-241). New York: Plenum Press.

Reid, D. K. (1981). Learning and development from a Piagetian perspective: The exceptional child. In J. Sigil, D. Brodzinsky, & R. Golinkoff (Eds.), New directions in Piagetian hheory and practice (pp. 339-344). Hillsdale, New Jersey: Earlbaum.

Reid, D. K., & Knight-Arest, I. (1979, July). Cognitive processing in learning disabled and normally achieving boys in a goal-oriented task. Paper presented at the NATO International Conference on Intelligence and Learning, York, England.

Reid, D. K., Knight-Arrest, I., & Hresko, W. P. (1981). Cognitive development in learning disabled children. In J. Gottlieb & S. S. Strichart (Eds.), Developmental theories and research in learning disabilities (pp. 169-212). Baltimore: University Park Press.

Reid, J. B. (1970). Reliability assessment of observation data: A possible methodological problem. Child Development, 41, 1143-1150.

Richman, L. C., & Lindgren, T. D. (1980). Patterns of intellectual ability in children with verbal deficits. Journal of Abnormal Child Psychology, 8, 65-81.

Romanczyk, R. G., Kent, R. W., Diament, C., & O'Leary, K. D. (1971, May). Measuring the reliability of observational data: A reactive process. Lawrence, KS. Paper presented at the Second Annual Symposium on Behavioral Analysis.

Ross, R. J., Hubbell, C., Ross, C. G., & Thompson, M. B. (1976). The training and transfer of formal thinking tasks in college students. Genetic Psychology Monographs, 93, 171-187.

Rourke, B. P., & Finlayson, M. A. (1978). Neuropsychological significance of variations as patterns of academic performance: Verbal and visual-spatial abilities. Journal of Abnormal Child Psychology, 6, 121-133.

Rourke, B. P., Dietrich, D. M., & Young, G. C. (1973). Significance of WISC verbal-performance discrepancies for younger children with learning disabilities. Perceptual and Motor Skills, 36, 275-282.

Rourke, B. P., & Telegdy, G. A. (1971). Lateralizing significance of WISC verbal-performance discrepancies for older children with learning disabilities. Perceptual and Motor Skills, 33, 875-883.

Rourke, B. P., Young, G. C., & Flewelling, R. W. (1971). The relationship between WISC verbal-performance discrepancies and selected verbal, auditory-perceptual, visual-perceptual, and problem solving abilities & children with learning disabilities. Journal of Clinical Psychology, 27, 475-479.

Schiff, M. M., Kaufman, A. S., & Kaufman, N. L. (1981). Scatter analysis of WISC-R profiles for learning disabled children with superior intelligence. Journal of Learning Disabilities, 14, 400-404.

Schmid-Kitsikis, E. (1969). L'examen des operations de l'intelligence: Psychopathologie de l'enfant [Study of process of intelligence: Child psychopathology]. Neuchatel: Delachaux et Niestle.

Schmid-Kitsikis, E. (1972). Exploratory studies in cognitive development. In F. J. Monks, W. W. Hartup, & J. deWit (Eds.). Determinants of behavioral development (pp. 51-63). New York: Academic Press.

Schmid-Kitsikis, E. (1973). Piagetian theory and its approach to psychopathology. American Journal of Mental Deficiency, 77, 694-705.

Schmid-Kitsikis, E. (1976). The cognitive mechanisms underlying problem solving in psychotic and mentally retarded children. In B. Inhelder and H. Chipman (Eds.), Piaget and his school (pp. 234-355). New York: Springer-Verlag.

Schoonover, I. M., & Hertel, R. K. (1970). Diagnostic implications of WISC scores. Psychological Reports, 26, 967-973.

Seashore, H. G. (1951). Differences between verbal and performance IQs on the Wechsler Intelligence Scale for Children. Journal of Consulting Psychology, 15, 62-67.

Shayer, M. (1979). Has Piaget's construct of formal operational thinking any utility? British Journal of Educational Psychology, 49, 265-276.

Siegel, E. (1974). The exceptional child grows up. New York: E. P. Dutton.

Siegler, R. S. (1981). Developmental sequences within and between concepts. Monographs of the Society for Research in Child Development, 46(2, Serial No. 189).

Siegler, R. S., Liebert, D. E., & Liebert, R. M. (1973). Inhelder and Piaget's pendulum problem: Teaching preadolescents to act as scientists. Developmental Psychology, 9, 97-101.

Sigel, I. E., McGillicuddy-Delisi, A. V., Flaugher, J., & Rock, D. A. (1983). Parents as teachers of their own learning disabled children (Research Rep. No. 83-21). Princeton, NJ: Educational Testing Service.

Silvius, J. (1974). A study of the comparative performance of learning-disabled and normal children on Piagetian tests of conservation. Unpublished doctoral dissertation, Northwestern University.

Skrtic, T. M. (1980). Formal reasoning abilities of learning disabled adolescents: Implications for mathematics instruction (Research Report No. 7). Lawrence, KS: University of Kansas, Institute for Research in Learning Disabilities.

Somerville, S. C., (1974). The pendulum problem: Patterns of performance defining developmental stages. British Journal of Educational Psychology, 44, 430-443.

Stephens, B. (1977). Piagetian theory-applications for the mentally retarded and visually handicapped. In J. F. Magary, M. K. Poulsen, P. J. Levinson, & P. A. Taylor (Eds.), Piagetian theory and the helping professions. Los Angeles: University of Southern California.

Stephens, B. and McLaughlin, J. (1974). Two year gains in reasoning by retarded and nonretarded persons. American Journal of Mental Deficiency, 1974, 79, 116-126.

Sternlight, M. (1981). The development of cognitive judgment in the mentally retarded: A selective review of Piagetian inspired research. The Journal of Genetic Psychology, 139, 55-68.

Stone, C. A. (1977). Logical competence and psychological processing accounts of the transition from concrete to formal operation. Dissertation Abstracts International, 38, 939b.

Stone, C. A. (1980). Adolescent cognitive development: Implications for learning disabilities. Bulletin of the Orton Society, 30, 79-93.

Stone, C. A. (1981). Reasoning disorders in learning-disabled adolescents. The Exceptional Child, 28, 43-53.

Stone, C. A., & Day, M. C. (1978). Levels of availability of a formal operational strategy. Child Development, 49, 1054-1065.

Stone, C. A., & Day, M. C. (1980). Competence and performance models and the characterization of formal operational skills. Human Development, 23, 323-353.

Strauss, A. A., & Kephart, N. C. (1955). Psychpathology and education of the brain-injured child: Vol. 2. Progress in theory and clinic. New York: Grune & Stratton.

Strauss, A. A., & Werner, H. (1942). Disorders of conceptual thinking in the brain-injured child. Journal of Nervous and Mental Disease, 96, 153.

Taplin, P. S., & Reid, J. B. (1973). Effects of instructional set and experimenter influence on observer reliability. Child Development, 44(3), 547-554.

Tarver, S. G., Hallahan, D. P., Kauffman, J. M., & Ball, D. W. (1976). Verbal rehearsal and selective attention in children with learning disabilities: A developmental lag. Journal of Experimental Child Psychology, 22, 375-385.

Tarver, S. G., & Maggiore, R. (1979). Cognitive development in learning disabled boys. Learning Disability Quarterly, 2, 78-84.

Tatsuoka, M. M. (1971). Multivariate Analysis: Techniques for Educational and Psychological Research. New York: Wiley.

Torgesen, J. K., & Licht, B. G. (1983). The learning disabled child as an inactive learner: Retrospect and prospects. In J. C. McKenney & L. Feagans (Eds.), Topics in learning disabilities, Vol.1. Norwood, NJ: Ablex Press.

Vance, H. B., Singer, M. C., & Engin, A. W. (1980). WISC-R subtest differences for male and female LD children and youth. Journal of Clinical Psychology, 36, 953-957.

Waryas, C. L., & Crowe, T. A. (1982). Language delay. In N. J. Lass, L. V. McReynolds, I. L. Northern, & D. E. Yoder (Eds.), Speech, language, and hearing: Vol. 2. Pathologies of speech and language (pp. 761-779). Philadelphia: W. B. Saunders.

Wechsler, D. (1939). The Measurement of adult intelligence (1st ed.). Baltimore: Williams and Wilkins.

Wechsler, D., & Jaros, E. (1965). Schizophrenic patterns on the WISC. Journal of Clinical Psychology, 21, 288-292.

Weisz, J. R., & Yeates, K. O. (1981). Cognitive development in retarded and nonretarded persons: Piagetian tests of the similar structure hypothesis. Psychological Bulletin, 90, 153-178.

Weisz, J. R., & Zigler, E. (1979). Cognitive development in retarded and nonretarded persons: Piagetian tests of the similar sequence hypothesis. Psychological Bulletin, 86, 831-851.

Wener, B. D., & Templer, D. I. (1976). Relationship between WISC verbal-performance discrepancies and motor and psychomotor abilities in children with learning disabilites. Perceptual and Motor Skills, 42, 125-126.

Wilcox, E. (1970). Identifying characteristics of the NH adolescent. In L. E. Anderson (Ed.), Helping the adolescent with the hidden handicap (pp. 5-12). Belmont, CA: Fearon.

Wilton, K. M., & Boersma, F. J. (1974). Conservation research with the mentally retarded. In M. R. Ellis, International Review of Research in Mental Retardation (Vol. 7, pp. 113-144). New York: Academic Press.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Wollman, W. (1977). Controlling variables: A neo-Piagetian developmental sequence. Science Education, 61, 385-391.

Wong, B. Y. L. (1979). Increasing retention of main ideas through questioning strategies. Learning Disability Quarterly, 2, 42-47.

Woodcock, R. W., & Johnson, M. B. (1977). Woodcock-Johnson psychoeducational battery. Boston: Teaching Resources.

Zimmerman, I. L., & Woo-Sam, J. (1972). Research with the Wechsler intelligence scale for children: 1969-1970. Psychology in the Schools, 9, 232-271.

Table 1
Group Composition by Subject

Subject #	Sex	WISC-R IQ Scores VIQ	PIQ	FSIQ	WISC-R Subtest Scores Voc.	Comp.	BD	OA
		Group 1 No Discrepancy LD						
101	M	100	100	100	8	13	8	-
104	F	87	88	87	7	8	7	6
301	M	100	106	102	8	10	11	-
302	M	101	100	101	10	9	14	7
402	M	106	106	106	9	16	9	12
404	M	102	108	105	9	12	12	13
601	M	107	98	102	12	13	11	10
602	M	92	91	91	14	7	7	10
603	M	94	95	93	8	11	12	9
643	F	96	96	96	9	11	9	8
		Group 3 Low Performance LD						
610	M	108	72	89	13	11	4	7
631	F	103	85	93	-	11	8	-
633	M	115	86	101	11	12	11	5
822	F	108	81	94	9	13	6	7
824	M	114	95	105	14	9	9	7
826	M	115	92	105	11	15	9	9*
833	M	105	74	89	12	11	3	2
		Group 5 Low Normal						
119	F	88	98	92	8	7	13	11
124	M	92	90	90	7	9	8	11
228	M	92	102	96	7	9	10	9
427	M	94	104	98	8	6	12	7
650	M	97	98	97	8	8	8	7
871	M	87	91	88	8	9	5	11

* S included even though BD and OA higher than 8.

Subject #	Sex	WISC-R IQ Scores VIQ	PIQ	FSIQ	WISC-R Subtest Scores Voc.	Comp.	BD	OA
		Group 2 Low Verbal LD						
102	M	87	114	100	8	8	15	17
501	F	77	96	85	6	7	10	7
503	M	87	102	93	8	6	10	12
504	M	85	111	97	6	8	14	13
505	M	80	106	91	6	8	13	17
509	M	97	120	108	10	8	14	13
607	M	96	120	107	8	11	15	12
635	M	103	120	112	8	7	14	14
821	M	91	120	104	7	11	13	19
832	M	85	120	101	6	9	12	19
		Group 4 High Normal						
212	F	101	108	104	8	-	7	11
217	F	102	106	104	9	11	12	11
418	M	105	104	104	8	12	11	10
421	M	107	112	110	10	11	12	13
422	M	106	115	111	9	11	12	13
423	M	100	106	102	7	15	10	9
425	M	97	105	101	10	10	10	13
834	F	115	117	118	10	15	11	14
835	M	102	101	101	8	9	11	9
836	F	100	106	102	10	9	13	10
		Group 6 Fourth Grade						
708	M	102	100	101	9	12	13	11
810	M	105	102	103	11	11	11	10
811	F	102	96	100	14	8	11	7
841	M	107	98	102	10	12	11	8
901	M	102	96	100	12	8	9	13
903	M	107	114	111	12	12	11	14
905	F	97	105	101	10	12	9	8
906	F	105	109	106	13	11	10	13
907	M	92	91	91	7	13	7	9
908	M	101	91	96	9	14	12	10

Table 2

Summary of Group Composition

	N	Sex	VIQ	PIQ	FSIQ
Group 1: ND No discrepancy LD	10	M = 8 F = 2	97.5 (5.5) range 87-107	99.1 (6.6) range 88-108	98.0 (6.2) range 87-106
Group 2: LV Low verbal LD	10	M = 9 F = 1	88.8 (8.0) range 77-103	112.9 (8.9) range 96-120	99.8 (8.4) range 85-112
Group 3: LP Low performance LD	7	M = 5 F = 2	109.7 (5.0) range 103-115	83.6 (8.6) range 72- 95	96.6 (7.0) range 89-105
Group 4: High High normal	10	M = 6 F = 4	103.5 (5.1) range 97-115	108.0 (5.1) range 101-117	105.7 (5.6) range 101-118
Group 5: Low Low normal	6	M = 5 F = 1	91.7 (3.7) range 87- 97	97.2 (5.7) range 90-104	93.5 (4.1) range 88- 98
Group 6: N9 High & Low Combined	16	M = 11 F = 5	99.1 (7.4) range 87-115	104.0 (7.4) range 90-117	101.1 (7.8) range 88-118
Group 7: N4 Fourth Grade	10	M = 7 F = 3	102.0 (4.6) range 92-107	100.2 (7.5) range 91-114	101.1 (5.3) range 91-111

Note. Numbers in parentheses are standard deviations.

Table 3

Major Difficulty Cited for Each Subject in the Three Groups

NO DISCREPANCY	LOW VERBAL	LOW PERFORMANCE
S1 Reading	S1 Reading	S1 Hyperactivity
S2 Conceptualization	S2 Abstract Reasoning	S2 General Coordination
S3 Revisualization	S3 Reading	S3 Gross Motor Skills
S4 Concept Learning	S4 Reading Comprehension	S4 Visual-Spatial Orientation
S5 General Organization	S5 Syntax	S5 Thinking/Reasoning
S6 Main Idea		S6 Visual Motor Integration
S7 Abstract Reasoning		
S8 Reading Comprehension		
S9 Organization		

Table 4

Percent of Each LD Group Labelled by Teacher as Having a Specific Problem

	NO DISCREPANCY	LOW VERBAL	LOW PERFORMANCE
N =	9	5	6
Language Problem?	56	60	17
Reading Problem?	100	80	33
Nonverbal Problem?	44	20	100
Math Problem?	89	60	83
Thinking Skill Problem?	100	40	50

Table 5

Percent of Subjects in Each Group

Rated Moderately or Severely Impaired in Each Area

	NO DISCREPANCY	LOW VERBAL	LOW PERFORMANCE
Language			
Verbal Memory	56	40	17
Comprehension	89	80	17
Expressive	56	60	17
Reading			
Decoding	44	80	0
Comprehension	100	80	17
Nonverbal Problem			
Visual Motor Integration	22	20	83
Figure-Ground	11	0	33
Orientation	22	0	67
Social Perception	0	0	67
Math Problems			
Computation	33	40	33
Concepts	44	40	50
Word Problems	89	60	67
Thinking Skills			
Organization and Study Skills	78	40	67
Concept Learning	44	40	50
Logical Reasoning	44	40	50

Table 6

Percentage Agreement per Case on Rods Preliminary Coding Sheet
Between a Criterion (A.S.) and Three Coders (F.M., C.A. & L.C.)

Coders	Case 1	Case 2	Case 3	Case 4	Case 5	Average
F.M.	.73	.78	.83	.89	.72	.79
C.A.	.88	.72	.75	.80	.74	.78
L.C.	.77	.71	.76	.71	.82	.75

Table 7

Percentage Agreement per Case on Conductivity Preliminary Coding Sheet
Between Two Coders (C.A. & F.M.)

Case 1	Case 2	Case 3	Case 4	Case 5	Average
.72	.76	.85	.89	.78	.80

Table 8

Intraclass Correlations for Preliminary Coding Sheet Summary Codes
Between a Criterion (A.S.) and Three Coders (C.A., L.C. & F.M.)

Coders	Total # of Tests	Total # of Unconf.Tests	# Unconf.Tests Before Summary	# Unconf. Tests (Resp.to Probes)	Total # Unconf.Var.
C.A.	.79	.99	.94	.87	.83
L.C.	.97	.98	.92	.99	.75
F.M.	.83	.99	.83	.98	.96

Table 9

Intraclass Correlations for Non-verbal Summary Codes
Between Two Coders (A.S. & F.M.)

Code	Intraclass Correlation
Number of uncredited unconfounded rod pairs during Test Construction Phase 1 (TC1)	.90
Number of uncredited unconfounded rod pairs during Test Construction Phase 2 (TC2)	.79
Number of tests with counteracting variables during TC1	.86
Number of tests with counteracting variables during TC2	.75

Table 10

Intraclass Correlations in Rod Scanning and Selecting Codes
for Two Coders (F.M. and C.A.)

Code	Intraclass Correlation
Initial Rod Selection Time; minimum	1.00
Initial Rod Selection Time; maximum	.40
Initial Rod Selection Time; median	- .25
Selection Position; Number Adjacent Rods Selected	1.00
Selection Position; Selected Rods w/L--R Consistency	1.00
Selection Time Between Rods; minimum	1.00
Selection Time Between Rods; maximum	.77
Selection Time Between Rods; median	1.00
Number of Simultaneous Selections	.43
Number of False Starts	.47
Number of Rods Separated When Discarded	.75
Number of Array Driven Comments	1.00

Table 11

Intraclass Correlations on Articulation Code
Between Two Coders (L.C. & F.M.)

Code	Intraclass Correlation
Proportion of non-zero scores*	.93
Proportion of 1s	1.00
Proportion of 2s	.64
Proportion of 3s	.68
Proportion of 4s	.97
Proportion of 5s	.88

Table 12

Intraclass Correlations on Generality Code
Between Two Coders (C.A. and F.M.)

Code	Intraclass Correlation
Proportion of non-zero scores*	.98
Proportion of 1s	.97
Proportion of 2s	.05
Proportion of 3s	.88

Table 13

Intraclass Correlations on Explanatory Principle Code
Between Two Coders (C.A. and F.M.)

Code	Intraclass Correlation
Proportion of non-zero scores*	.95
Proportion of 1s	.99
Proportion of 2s	.64
Proportion of 3s	.93
Proportion of 4s	1.00

* A score of zero was given to tests for which this code was not applicable.

Table 14

Intraclass Correlations on Prediction Intention Code
Between Two Coders (F.M. and S.W.)

Code	Intraclass Correlation
Proportion of non-zero scores*	1.00
Proportion of 1s	1.00
Proportion of 2s	.99
Proportion of 3s	.82
Proportion of 4s	.98
Proportion of 5s	1.00

Table 15

Intraclass Correlations on Prediction Explicitness Code
Between Two Coders (F.M. and S.W.)

Code	Intraclass Correlation
Proportion of non-zero scores*	.71
Proportion of 1s	.97
Proportion of 2s	.91
Proportion of 3s	.75

* A score of zero was given to tests for which this code was not applicable.

Table 16

Intraclass Correlations on Attribute Code
Between Two Coders (L.C. and D.M.)

Code	Intraclass Correlation
Proportion of 1s	.97
Proportion of 2s	.89
Proportion of 3s	.95
Proportion of 4s	1.00
Proportion of 5s	1.00
Proportion of 6s	1.00
Proportion of 7s	.92
Proportion of 8s	.89

Table 17

Comparison of ND, LV, and LP Groups:

Summary of Analyses of Variance and Covariance on

Proportion of Unconfounded Tests

Test Phase and Source	df	SS	MS	F	p
TC1:					
Group	2	.091	.045	1.12	.34
Error	24	.972	.040		
TC2:					
Group	2	.088	.044	1.12	.33
Error	24	.910	.038		
TC2 (TC1 Covaried):					
Constant (=Improvement)	1	.732	.732	22.47	.001
Regression[a]	1	.161	.161	4.93	.036
Group	2	.034	.017	.53	.598
Error	23	.749	.033		

[a] A test for lack of parallelism of regression slopes was nonsignificant.

Table 18

Mean Proportion of Unconfounded Tests for Each Group

During Each Test Construction (TC) Phase

		TC1	TC2	
Group	N		Raw Scores	Adjusted for Initial Performance
ND	10	.22 (.16)	.35 (.23)	.38
LV	10	.34 (.24)	.45 (.17)	.43
LP	7	.33 (.20)	.49 (.16)	.48
Total	27	.29 (.20)	.42 (.20)	.42

Note. Numbers in parentheses are standard deviations. SPSS Manova does not provide this parameter for adjusted scores.

Table 19

Mean Number Correct on Test Choice and Test Evaluation Phases

		Question Type	
Group	N	Choice[a]	Evaluation[b]
ND	10	1.80 (1.62)	2.40 (1.84)
LV	10	2.60 (1.90)	1.90 (1.37)
LP	7	3.29 (1.70)	4.00 (1.53)

Note. Numbers in parentheses are standard deviations.

[a]Possible score = 5.0.

[b]Possible score = 6.0.

Table 20

Effect Sizes for Comparisons of Group Means in Each Rods Test Phase

Comparison	Test Construction		Improvement	Test Choice	Test Eval.
	First	Second			
ND vs LV	-.62	-.51	-.24	- .45	.31
ND vs LP	-.63	-.72	-.52	-1.33	-1.17
LV vs LP	.06	-.21	-.28	- .38	-1.46

Table 21

Comparison of ND, N9, and N4 Groups:

Summary of Analyses of Variance and Covariance on

Proportion of Unconfounded Tests

Test Phase and Source	df	SS	MS	F	p
TC1:[a]					
Group	2	.249	.124	3.56	.040
Error	33	1.152	.035		
TC2:					
Group	2	.575	.288	5.51	.009
Error	33	1.721	.052		
TC2 (TC1 Covaried):					
Constant (=Improvement)	1	.492	.492	13.02	.001
Regression[b]	1	.511	.511	13.51	.001
Group	2	.165	.083	2.18	.129
Error	32	1.210	.038		

[a] A supplementary test (Cochran's C) indicated significant heterogeneity of variance. Because violations of the homogeneity of variance assumption become important with unequal cell sizes, a Kruskal-Wallis analysis of variance was also carried out. Results were comparable.

[b] A test for lack of parallelism of regression slopes was nonsignificant.

Table 22

Mean Proportion of Unconfounded Tests for Each Group

During Each Test Construction (TC) Phase

		TC1	TC2	
Group	N		Raw Scores	Adjusted for Initial Performance
ND	10	.22 (.16)	.35 (.23)	.38
N9	16	.38 (.24)	.55 (.28)	.47
N4	10	.20 (.08)	.25 (.11)	.29
Total	36	.28 (.20)	.41 (.26)	.40

Note. Numbers in parentheses are standard deviations. SPSS Manova does not provide this parameter for adjusted scores.

Table 23

Mean Number Correct on Test Choice and Test Evaluation Phases

		Question Type	
Group	N	Choice[a]	Evaluation[b]
ND	10	1.80 (1.62)	2.40 (1.84)
N9	16	3.12 (1.86)	4.12 (1.36)
N4	10	2.40 (1.78)	2.10 (1.60)

Note. Numbers in parentheses are standard deviations.

[a]Possible score = 5.0.

[b]Possible score = 6.0.

Table 24

Effect Sizes for Comparisons of Group Means in Each Rods Test Phase

Comparison	Test Construction		Improvement	Test Choice	Test Eval.
	First	Second			
ND vs N9	-.73	- .83	-.44	-.75	-1.11
ND vs N4	.16	.46	.47	-.35	.17
N9 vs N4	.89	1.29	.91	.42	1.39

Table 25

Comparison of LV, N9, and N4 Groups:

Summary of Analyses of Variance and Covariance on

Proportion of Unconfounded Tests

Test Phase and Source	df	SS	MS	F	p
TC1:					
Group	2	.20	.10	2.29	.117
Error	33	1.44	.04		
TC2:[a]					
Group	2	.53	.27	5.83	.007
Error	33	1.51	.05		
TC2 (TC1 Covaried):					
Constant (=Improvement)	1	.61	.61	19.54	.001
Regression[b]	1	.51	.51	16.16	.001
Group	2	.20	.10	3.12	.058
Error	32	1.01	.03		

[a] A supplementary test (Cochran's C) indicated significant heterogeneity of variance. Because violations of the homogeneity of variance assumption become important with unequal cell sizes, A Kruskal-Wallis analysis of variance was also carried out. Results were comparable.

[a] A test for lack of parallelism of regression slopes was nonsignificant.

Table 26

Mean Proportion of Unconfounded Tests for Each Group

During Each Test Construction (TC) Phase

		TC1	TC2	
Group	N		Raw Scores	Adjusted for Initial Performance
LV	10	.34 (.24)	.45 (.17)	.42
N9	16	.38 (.24)	.54 (.28)	.50
N4	10	.20 (.08)	.25 (.11)	.31
Total	36	.32 (.22)	.43 (.24)	.42

Note. Numbers in parentheses are standard deviations. SPSS Manova does not provide this parameter for adjusted scores.

Table 27

Mean Number Correct on Test Choice and Test Evaluation Phases

Group	N	Question Type	
		Choice[a]	Evaluation[b]
LV	10	2.60 (1.90)	1.90 (1.37)
N9	16	3.12 (1.86)	4.12 (1.36)
N4	10	2.40 (1.78)	2.10 (1.60)

Note. Numbers in parentheses are standard deviations.

[a]Possible score = 5.0.

[b]Possible score = 6.0.

Table 28

Effect Sizes for Comparisons of Group Means in Each Rods Test Phase

Comparison	Test Construction		Improvement	Test Choice	Test Eval.
	First	Second			
LV vs N9	-.13	-.40	-.43	- .28	-1.63
LV vs N4	.82	.89	.64	.11	- .13
N9 vs N4	.89	1.30	1.07	.40	1.39

Table 29

Comparisons of LP, N9, and N4 Groups:

Summary of Analyses of Variance and Covariance on

Proportion of Unconfounded Tests

Test Phase and Source	df	SS	MS	F	p
TC1:					
Group	2	.19	.10	2.46	.102
Error	30	1.18	.04		
TC2:[a]					
Group	2	.55	.28	5.90	.007
Error	30	1.41	.05		
TC2 (TC1 Covaried):					
Constant (=Improvement)	1	.53	.53	16.34	.001
Regression[b]	1	.47	.47	14.66	.001
Group	2	.19	.09	2.91	.070
Error	29	.93	.03		

[a] A supplementary test (Cochran's C) indicated significant heterogeneity of variance. Because violations of the homogeneity of variance assumption become important with unequal cell sizes, A Kruskal-Wallis analysis of variance was also carried out. Results were comparable.

[a] A test for lack of parallelism of regression slopes was nonsignificant.

Table 30

Mean Proportion of Unconfounded Tests for Each Group

During Each Test Construction (TC) Phase

		TC1	TC2	
Group	N		Raw Scores	Adjusted for Initial Performance
LP	7	.33 (.20)	.49 (.16)	.47
N9	16	.38 (.29)	.54 (.28)	.49
N4	10	.20 (.08)	.25 (.11)	.31
Total	33	.31 (.21)	.44 (.25)	.43

Note. Numbers in parentheses are standard deviations. SPSS Manova does not provide this parameter for adjusted scores.

Table 31

Mean Number Correct on Test Choice and Test Evaluation Phases

Group	N	Question Type: Choice[a]	Question Type: Evaluation[b]
LP	7	3.29 (1.70)	4.00 (1.53)
N9	16	3.12 (1.86)	4.12 (1.36)
N4	10	2.40 (1.78)	2.10 (1.60)

Note. Numbers in parentheses are standard deviations.

[a]Possible score = 5.0.

[b]Possible score = 6.0.

Table 32

Effect Sizes for Comparisons of Group Means in Each Rods Test Phase

	Test Construction				
Comparison	First	Second	Improvement	Test Choice	Test Eval.
LP vs N9	-.20	-.22	-.13	.09	- .09
LP vs N4	.92	1.07	.89	.51	1.21
N9 vs N4	.89	1.30	1.02	.40	1.39

Table 33

Comparison of ND, High IQ, and Low IQ Groups:
Summary of Analyses of Variance and Covariance on
Proportion of Unconfounded Tests

Test Phase and Source	df	SS	MS	F	p
TC1:					
Group	2	.38	.19	4.96	.016
Error	23	.87	.04		
TC2:					
Group	2	.68	.34	6.72	.005
Error	23	1.16	.05		
TC2 (TC1 Covaried):					
Constant (=Improvement)	1	.68	.68	15.40	.001
Regression[a]	1	.19	.10	4.36	.049
Group	2	.20	.10	2.31	.123
Error	22	.97	.04		

[a] A test for lack of parallelism of regression slopes was nonsignificant.

Table 34

Mean Proportion of Unconfounded Tests for Each Group During Each Test Construction (TC) Phase

		TC1	TC2	
Group	N		Raw Scores	Adjusted for Initial Performance
ND	10	.22 (.16)	.35 (.23)	.39
High	10	.47 (.22)	.67 (.17)	.60
Low	6	.22 (.21)	.32 (.29)	.36
Total	26	.31 (.22)	.47 (.27)	.46

Note. Numbers in parentheses are standard deviations. SPSS Manova does not provide this parameter for adjusted scores.

Table 35

Mean Number Correct on Test Choice and Test Evaluation Phases

		Question Type	
Group	N	Choice[a]	Evaluation[b]
ND	10	1.80 (1.62)	2.40 (1.84)
High	10	4.30 (.95)	4.10 (1.08)
Low	6	1.17 (.48)	3.67 (1.75)

Note. Numbers in parentheses are standard deviations.

[a]Possible score = 5.0.

[b]Possible score = 6.0.

Table 36

Effect Sizes for Comparisons of Group Means in Each Rods Test Phase

Comparison	Test Construction		Improvement	Test Choice	Test Eval.
	First	Second			
ND vs High	-1.30	-1.43	-.97	-1.88	-1.13
ND vs Low	- .02	.13	.14	.47	- .70
High vs Low	1.12	1.56	1.12	3.85	.32

Table 37

Comparison of LV, High IQ, and Low IQ Groups:

Summary of Analyses of Variance and Covariance on

Proportion of Unconfounded Tests

Test Phase and Source	df	SS	MS	F	p
TC1:					
Group	2	.23	.12	2.28	.125
Error	23	1.17	.05		
TC2:					
Group	2	.51	.26	6.18	.007
Error	23	.95	.04		
TC2 (TC1 Covaried):					
Constant (=Improvement)	1	.81	.81	23.97	.001
Regression[a]	1	.21	.21	6.34	.02
Group	2	.22	.11	3.32	.055
Error	22	.74	.03		

[a] A test for lack of parallelism of regression slopes was significant. Therefore, these results should be interpreted with caution.

Table 38

Mean Proportion of Unconfounded Tests for Each Group

During Each Test Construction (TC) Phase

		TC1	TC2	
Group	N		Raw Scores	Adjusted for Initial Performance
LV	10	.34 (.24)	.45 (.17)	.45
High	10	.47 (.22)	.67 (.17)	.62
Low	6	.22 (.21)	.32 (.29)	.38
Total	26	.36 (.24)	.50 (.24)	.50

Note. Numbers in parentheses are standard deviations. SPSS Manova does not provide this parameter for adjusted scores.

Table 39

Mean Number Correct on Test Choice and Test Evaluation Phases

		Question Type	
Group	N	Choice[a]	Evaluation[b]
LV	10	2.60 (1.90)	1.90 (1.37)
High	10	4.30 (.95)	4.40 (1.08)
Low	6	1.17 (.48)	3.67 (1.75)

Note. Numbers in parentheses are standard deviations.

[a]Possible score = 5.0.

[b]Possible score = 6.0.

Table 40

Effect Sizes for Comparisons of Group Means in Each Rods Test Phase

	Test Construction				
Comparison	First	Second	Improvement	Test Choice	Test Eval.
LV vs High	- .54	-1.09	- .94	-1.13	-2.03
LV vs Low	.53	.64	.38	.92	-1.16
High vs Low	1.13	1.73	1.32	3.85	.54

Table 41

Comparison of LP, High IQ, and Low IQ Groups:

Summary of Analyses of Variance and Covariance on

Proportion of Unconfounded Tests

Test Phase and Source	df	SS	MS	F	p
TC1:					
Group	2	.24	.12	2.61	.099
Error	20	.90	.05		
TC2:					
Group	2	.47	.23	5.54	.012
Error	20	.85	.04		
TC2 (TC1 Covaried):					
Constant (=Improvement)	1	.73	.73	20.37	.001
Regression[a]	1	.17	.17	4.69	.043
Group	2	.18	.09	2.52	.107
Error	19	.68	.04		

[a] A test for lack of parallelism of regression slopes was significant. Therefore, these results should be interpreted with caution.

Table 42

Mean Proportion of Unconfounded Tests for Each Group

During Each Test Construction (TC) Phase

		TC1	TC2	
Group	N		Raw Scores	Adjusted for Initial Performance
LP	7	.33 (.20)	.49 (.16)	.50
High	10	.47 (.22)	.67 (.17)	.62
Low	6	.22 (.21)	.32 (.29)	.37
Total	23	.36 (.23)	.53 (.24)	.52

Note. Numbers in parentheses are standard deviations. SPSS Manova does not provide this parameter for adjusted scores.

Table 43

Mean Number Correct on Test Choice and Test Evaluation Phases

		Question Type	
Group	N	Choice[a]	Evaluation[b]
LP	7	3.29 (1.70)	4.00 (1.53)
High	10	4.30 (.95)	4.40 (1.08)
Low	6	1.17 (.48)	3.67 (1.75)

Note. Numbers in parentheses are standard deviations.

[a]Possible score = 5.0.

[b]Possible score = 6.0.

Table 44

Effect Sizes for Comparisons of Group Means in Each Rods Test Phase

	Test Construction				
Comparison	First	Second	Improvement	Test Choice	Test Eval.
LP vs High	- .64	- .88	- .64	- .78	-.31
LP vs Low	.52	.83	.64	1.63	.20
High vs Low	1.13	1.71	1.28	3.85	.54

Table 45

Comparison of the Rods and Conductivity Tasks:

Mean Proportion Unconfounded Tests For Each Group

		Rods Task		
Group	N	TC1	TC2	Conductivity Task
ND	10	.22 (.16)	.35 (.23)	.31 (.20)
LV	10	.34 (.24)	.45 (.17)	.41 (.23)
LP	7	.33 (.20)	.49 (.16)	.46 (.18)
N9	16	.38 (.24)	.54 (.28)	.51 (.25)
N4	4	.20 (.08)	.25 (.11)	.31 (.18)
Total	53	.30	.43	.41

Note: Numbers in parentheses are standard deviations.

Table 46

Comparison of the Rods and Conductivity Tasks:

Results of Statistical Tests of Group Differences in Unconfounded Testing

Group Comparison	Rods Task[a]		Conductivity Task
	TC1	TC2	
ND vs LV vs LP	n.s.	n.s.	n.s.[b]
ND vs N9 vs N4	sig.	sig.	sig.[c]
LV vs N9 vs N4	n.s.	sig.	n.s.[d]
LP vs N9 vs N4	n.s.	sig.	n.s.[e]

Note. n.s = not significant (p .05); sig. = significant (p .05).

[a] F values for rods task analyses are presented elsewhere.

[b] F (2,24) = 1.30, p > .10.

[c] F (2,33) = 3.83, p < .05.

[d] F (2,33) = 2.47, p = .10.

[e] F (2,30) = 2.70, p = .08.

Table 47

Comparison of the Rods and Conductivity Tasks:

Effect Sizes For Group Comparisons of Unconfounded Testing

Comparison	Rods Task		Conductivity Task
	TC1	TC2	
ND vs LV	- .62	- .51	- .49
ND vs LP	- .63	- .72	- .77
LV vs LP	+ .06	- .20	- .28
ND vs N9	- .73	- .83	- .93
ND vs N4	+ .16	+ .46	- .01
N9 vs N4	+ .89	+1.29	+ .92
LV vs N9	- .13	- .42	- .45
LV vs N4	+ .82	+ .93	+ .43
N9 vs N4	+ .89	+1.35	+ .88
LP vs N9	- .20	- .23	- .20
LP vs N4	+ .92	+1.11	+ .72
N9 vs N4	+ .89	+1.34	+ .92

Table 48

Comparison of the Rods and Conductivity Tasks:

Proportion of Unconfounded Tests for Subjects in 3 Rods Strategy-Status Groups

Rods Strategy Status Group	N	Rods Task		Conductivity Task[a]
		TC1	TC2	
Absent	31	.18 (.11)	.28 (.14)	.33 (.23)
Elicitable	14	.35 (.13)	.64 (.15)	.50 (.15)
Spontaneous	8	.66 (.13)	.61 (.19)	.55 (.21)

Note. Criteria for assignment to strategy-status groups are described in the text.

[a] $\underline{F}$ (2,50) = 5.44, $\underline{p} < .01$.

Table 49

Comparison of Rods and Conductivity Tasks:

Correlations of Unconfounded Testing Scores

	Rods Task TC1	Rods Task TC2	Conductivity Task
TC1		.56[a]	.30[b]
TC2	.59[a]		.47[a]
Conductivity	.34[a]	.49[a]	

Note. Entries below the diagonal are zero-order Pearson correlations. Entries above the diagonal are partial correlations controlling for (WISC-R) fullscale IQ.

[a] $p < .01$.

[b] $p < .05$.

Table 50

Correlations of Woodcock-Johnson Concept Formation Standard Scores

With Rods and Conductivity Summary Scores

	Fullscale IQ	Prop.Unconf. Tests TC1	Test Choice	Prop.Unconf. Tests TC2	Test Eval.	Conduc.Prop. Unconf.Tests
Raw Correlations	.58[b]	.28[a]	.52[b]	.39[b]	.44[b]	.46[b]
Fullscale IQ Covaried	----	.11	.32[a]	.26[a]	.37[b]	.41[b]

Note. Correlations were calculated using all 43 ninth grade subjects. For the partial correlations,

df = 50.

[a] $p < .05$.

[b] $p < .01$.

Table 51

Mean Proportion of Reasoning and Noncompliance Codes by Group for Test Construction Phase 1

	Group				
Code	LV	LP	ND	N9	N4
High Articulation of Conclusion	62.4(24.6)	74.1(19.2)	62.6(27.9)	70.1(30.8)	52.7(28.0)
Lack of Explanatory Principles in Concl.	69.5(20.3)	64.4(29.0)	65.3(26.4)	56.6(38.2)	79.8(21.3)
Low Generality of Conclusions	71.1(28.6)	75.1(16.5)	57.3(31.7)	70.1(31.3)	89.2(17.6)
Inexplicitness of Predictions	43.6(35.6)	19.8(15.9	40.1(28.7)	30.0(34.5)	34.8(19.9)
High Degree of Intentionality in Predictions	19.9(23.0)	50.0(34.3)	33.7(29.9)	42.6(33.4)	45.9(31.7)
Noncompliance with Prompts	28.2(30.0)	21.6(26.7)	17.7(20.5)	8.8(14.8)	28.3(36.2)
N	10	7	10	14[a]	10

Note. Numbers in parentheses are standard deviations.

[a]Two subjects from the ninth grade control group were dropped due to missing data.

Table 52

Standardized Discriminant Function Coefficients[a] and

Group Centroids for Test Construction Phase 1

Variable	Standardized Coefficients	
	Function 1	Function 2
Articulation of conclusions	.141	.228
Use of explanatory principles in conclusions	.151	.629
Generality of conclusions	.284	.918
Explicitness of predictions	.514	-.290
Intentionality in predictions	-.757	.014
Noncompliance with prompts	.371	.494

Group	Centroids	
	Function 1	Function 2
ND	-.148	-.517
LV	.592	.205
LP	-.634	.446

[a] $N = 27$.

Table 53

Summary of Misclassification Analyses for Test Construction Phases 1 and 2

Test Construction Phase 1

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		ND	LV	LP	ND	LV	LP
	ND	6	4	0	2	6	2
Actual	LV	4	5	1	3	4	3
Group	LP	2	2	3	3	2	2

Test Construction Phase 2

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		ND	LV	LP	ND	LV	LP
	ND	6	3	1	1	5	4
Actual	LV	3	6	1	3	6	1
Group	LP	1	1	5	3	2	2

Table 54

Mean Proportion of Reasoning and Noncompliance Codes

by Group for Test Construction Phase 2

Code	Group				
	LV	LP	ND	N9	N4
High Articulation of Conclusion	60.6(29.2)	84.4(18.3)	73.2(27.4)	85.2(24.0)	64.9(21.8)
Lack of Explanatory Principles in Concl.	72.6(14.6)	56.1(36.8)	71.3(23.9)	63.8(42.2)	75.2(26.6)
Low Generality of Conclusions	59.5(38.0)	71.4(24.0)	65.7(33.9)	79.9(19.8)	85.9(23.2)
Inexplicitness of Predictions	32.4(24.7)	21.8(28.6)	22.3(35.1)	13.9(13.7)	25.2(16.8)
High Degree of Intentionality in Predictions	28.2(26.4)	46.4(49.9)	27.3(40.6)	33.5(33.2)	33.9(29.0)
Noncompliance with Prompts	21.2(20.9)	9.5(16.3)	23.9(31.6)	4.6(10.1)	18.7(21.4)
N	10	7	10	14[a]	10

Note. Numbers in parentheses are standard deviations.

[a]Two subjects from the ninth grade control group were dropped due to missing data.

Table 55

Standardized Discriminant Function Coefficients[a] and Group Centroids for Test Construction Phase 2

Variable	Standardized Coefficients	
	Function 1	Function 2
Articulation of conclusions	.910	-.357
Use of explanatory principles in conclusions	-.572	-.223
Generality of conclusions	.307	-.407
Explicitness of predictions	.093	.695
Intentionality in predictions	.446	.734
Noncompliance with prompts	-.453	-.233

Group	Centroids	
	Function 1	Function 2
ND	- .169	-.294
LV	- .603	.219
LP	1.103	.107

[a] $N = 27$.

Table 56

Standardized Discriminant Function Coefficients[a] and Group Centroids for Test Construction Phase 1

Variable	Standardized Coefficients	
	Function 1	Function 2
Articulation of conclusions	-.300	-.030
Use of explanatory principles in conclusions	.732	.036
Generality of conclusions	.616	.265
Explicitness of predictions	.156	.038
Intentionality in predictions	.288	.848
Noncompliance with prompts	.416	-.535

Group	Centroids	
	Function 1	Function 2
LV	-.050	-.668
N9	-.624	.300
N4	.924	.247

[a] N = 34. Two subjects from the ninth grade control group were dropped due to missing data.

Table 57

Summary of Misclassification Analyses for Test Construction Phases 1 and 2

Test Construction Phase 1

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		LV	N9	N4	LV	N9	N4
	LV	6	2	2	2	5	3
Actual	N9	1	11	2	3	7	4
Group	N4	2	2	6	4	3	3

Test Construction Phase 2

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		LV	N9	N4	LV	N9	N4
	LV	6	1	3	4	1	5
Actual	N9	2	11	1	2	10	2
Group	N4	2	3	5	4	6	0

Table 58

Standardized Discriminant Function Coefficients[a] and Group Centroids for Test Construction Phase 2

Variable	Standardized Coefficients	
	Function 1	Function 2
Articulation of conclusions	.498	.020
Use of explanatory principles in conclusions	-.085	.306
Generality of conclusions	.628	.893
Explicitness of predictions	-.324	.397
Intentionality in predictions	.067	.175
Noncompliance with prompts	-.609	.159

Group	Centroids	
	Function 1	Function 2
LV	-1.054	-.255
N9	.850	-.170
N4	-.136	.493

[a] $N = 34$. Two subjects from the ninth grade control group were dropped due to missing data.

Table 59

Standardized Discriminant Function Coefficients[a] and Group Centroids for Test Construction Phase 1

Variable	Standardized Coefficients	
	Function 1	Function 2
Articulation of conclusions	-.453	.390
Use of explanatory principles in conclusions	.775	.134
Generality of conclusions	.648	-.149
Explicitness of predictions	.354	-.705
Intentionality in predictions	.403	-.016
Noncompliance with prompts	.339	.660

Group	Centroids	
	Function 1	Function 2
LP	- .256	.463
N9	- .637	-.180
N	1.071	-.072

[a] N = 31. Two subjects from the ninth grade control group were dropped due to missing data.

Table 60

Summary of Misclassification Analyses for Test Construction Phases 1 and 2

Test Construction Phase 1

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		LP	N9	N4	LP	N9	N4
	LP	1	5	1	0	6	1
Actual	N9	1	10	3	3	7	4
Group	N4	0	2	8	1	3	6

Test Construction Phase 2

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		LP	N9	N4	LP	N9	N4
	LP	2	4	1	0	6	1
Actual	N9	1	11	2	4	8	2
Group	N4	1	4	5	1	5	4

Table 61

Standardized Discriminant Function Coefficients[a] and Group Centroids for Test Construction Phase 2

Variable		Standardized Coefficients	
		Function 1	Function 2
Articulation of conclusions	-.490	.348	
Use of explanatory principles in conclusions		.252	-.297
Generality of conclusions		.282	-.318
Explicitness of predictions	.185	.835	
Intentionality in predictions		.128	.717
Noncompliance with prompts	.617	.213	

Group	Centroids	
	Function 1	Function 2
LP	-.275	.591
N9	-.446	-.256
N4	.817	-.056

[a] N = 31. Two subjects from the ninth grade control group were dropped due to missing data.

Table 62

Standardized Discriminant Function Coefficients[a] and Group Centroids for Test Construction Phase 1

Variable		Standardized Coefficients	
		Function 1	Function 2
Articulation of conclusions	-.311	.181	
Use of explanatory principles in conclusions		.798	-.146
Generality of conclusions		.963	.519
Explicitness of predictions	.259	-.127	
Intentionality in predictions		.466	.273
Noncompliance with prompts	.232	-.545	

Group	Centroids	
	Function 1	Function 2
ND	- .566	-.367
N9	- .479	.275
N4	1.24	-.019

[a] N = 34. Two subjects from the ninth grade control group were dropped due to missing data.

Table 63

Summary of Misclassification Analyses for Test Construction Phases 1 and 2

Test Construction Phase 1

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		ND	N9	N4	ND	N9	N4
	ND	3	6	1	1	8	1
Actual	N9	4	8	2	5	5	4
Group	N4	0	2	8	0	3	7

Test Construction Phase 2

		Assigned Group					
		Standard Procedure			Jackknife Procedure		
		ND	N9	N4	ND	N9	N4
	ND	3	4	3	3	4	3
Actual	N9	0	12	2	2	10	2
Group	N4	1	4	5	3	5	2

Table 64

Standardized Discriminant Function Coefficients[a] and Group Centroids for Test Construction Phase 2

Variable	Standardized Coefficients	
	Function 1	Function 2
Articulation of conclusions	.634	-.199
Use of explanatory principles in conclusions	-.146	.318
Generality of conclusions	.349	.890
Explicitness of predictions	-.154	.409
Intentionality in predictions	.057	.371
Noncompliance with prompts	-.669	-.153

Group	Centroids	
	Function 1	Function 2
ND	-.534	-.440
N9	.650	-.062
N4	-.356	.527

[a] $N = 34$. Two subjects from the ninth grade control group were dropped due to missing data.

Table 65

Teacher Rating Scale: Percent Agreement With Master by Case for Each Category of Information

Variable	Subject Number 101	302	404	Average
rod selection	84	91	78	84
confounded/ unconfounded	96[a]	91[a]	100[a]	96[a]
reasons	44	68	22	45
variables mentioned	80	92	76	83
conclusions	64	59	47	57
explanations	56	53	53	54
Total	71	76	63	70

[a]These percentages ignore the distinction between U-[variable] and U-[variable] (See the TRS manual in Appendix C. This distinction was not picked up on by the raters and in fact occurs only three times in the three cases.

Table 66

Teacher Rating Scale: Percent Agreement With

Master by Case for Each Rater

	Subject Number			
	101	302	404	Across Subjects (180 responses)
Rater 1	74 (2)	72 (1)	60 (3)	69
Rater 2	66	84	58	69
Rater 3	69 (1)	74 (2)	70 (3)	71
Rater 4	74 (1)	76 (2)	67 (3)	72
Rater 5	71 (1)	71 (3)	58 (2)	67
Average	71	76	63	70

Note. Numbers in parentheses indicate the order in which each rater coded the three subjects. This information was unavailable for Rater 2.

Table 67

Teacher Rating Scale: Percent Agreement

With Master Coding -- Subject 101

	Rater					
Variable	R1	R2	R3	R4	R5	Ave.
rod selection	100	100	56	89	78	85
confounded/ unconfounded	100	89	100	100	89	96
reasons	67	56	22	33	44	44
variables mentioned	80	60	80	100	80	80
conclusions	67	33	89	56	78	65
explanations	33	56	67	67	56	56
Total (based on 50 responses)	74	66	69	74	71	71

Table 68

Teacher Rating Scale: Percent Agreement

With Master Coding -- Subject 302

	Rater					
Variable	R1	R2	R3	R4	R5	Ave.
rod selection	93	87	87	93	93	91
confounded/ unconfounded	93	100	100	80	80	91
reasons	73	93	47	80	47	68
variables mentioned	100	100	100	80	80	92
conclusions	27	60	60	60	87	59
explanations	47	73	47	60	40	53
Total (based on 80 responses)	72	84	74	76	71	76

Table 69

Teacher Rating Scale: Percent Agreement

With Master Coding -- Subject 404

	Rater					
Variable	R1	R2	R3	R4	R5	Ave.
rod selection	78	78	67	78	89	78
confounded/ unconfounded	100	100	100	100	100	100
reasons	22	22	22	33	11	22
variables mentioned	60	60	100	80	80	76
conclusions	22	67	44	33	67	47
explanations	78	22	89	78	0	53
Total (based on 50 responses)	60	58	70	67	58	63

Table 70

Mean Proportion of Reasoning and Unconfounded Testing Codes by Subject Across the Intervention Sessions

	Rods Pretest		Conductivity			Rods Posttest
	TC1	TC2	Baseline	Interven.	Baseline	TC3
Proportion of Specific Conclusions						
OP4	.43(14)	.77(13)	.40(5)	.32(18)[a]	.40(10)	.75(20)
OP5	.33(6)	.60(5)	.20(5)	.23(13)	.00(3)	.00(3)
OP9	.50(4)	.66(6)	.50(6)	.44(9)	.75(8)	.17(6)
Proportion of Highly Articulated Concl.						
OP4	.66(3)	.62(8)	.38(8)	.46(19)	.56(9)	.54(13)
OP5	1.00(6)	.60(5)	.83(6)	1.00(10)	1.00(3)	1.00(2)
OP9	1.00(4)	.75(4)	.38(8)	.50(8)	.50(8)	.75(4)
Prop. of Tests with Explan. Principles						
OP4	.00(3)	.33(6)	.33(6)	.62(19)	.00(13)	.00(13)
OP5	.85(7)	1.00(5)	1.00(5)	1.00(10)	1.00(3)	1.00(2)
OP9	.50(4)	.88(8)	1.00(9)	1.00(10	1.00(10)	.80(5)
Prop. of Predictions w/ High Intentionality						
OP4	.50(4)	1.00(9)	.29(17)	.44(18)	.53(19)	1.00(12)
OP5	.50(10)	.60(5)	1.00(5)	1.00(11)	1.00(3)	.50(4)
OP9	1.00(3)	1.00(5)	1.00(9)	.89(9)	.50(6)	.43(7)
Proportion of Highly Explicit Predictions						
OP4	.50(4)	.88(9)	.30(17)	.55(18)	.42(19)	.92(12)
OP5	.70(10)	.80(5)	1.00(6)	.73(11)	.33(3)	.75(4)
OP9	1.00(3)	.80(5)	1.00(9)	.66(9)	.33(6)	.43(7)
Prop. of Unconf.Tests Before Summary						
OP4	.00	.50	.18	.30	.42	.69
OP5	.27	.13	.50	.64	.33	.40
OP9	.22	.14	.60	.44	.33	.60
Number Unconfounded Variables						
OP4	0	2	1	1	2	5
OP5	1	2	1	2	0	5
OP9	2	0	0	1	0	5

Note. Numbers in parentheses are the number of nonzero scores in each category.

[a] Proportion scores for OP4 on C2 represent an average across both of his two intervention sessions.

Figure 1

Proportion of Unconfounded Tests
Before Summary

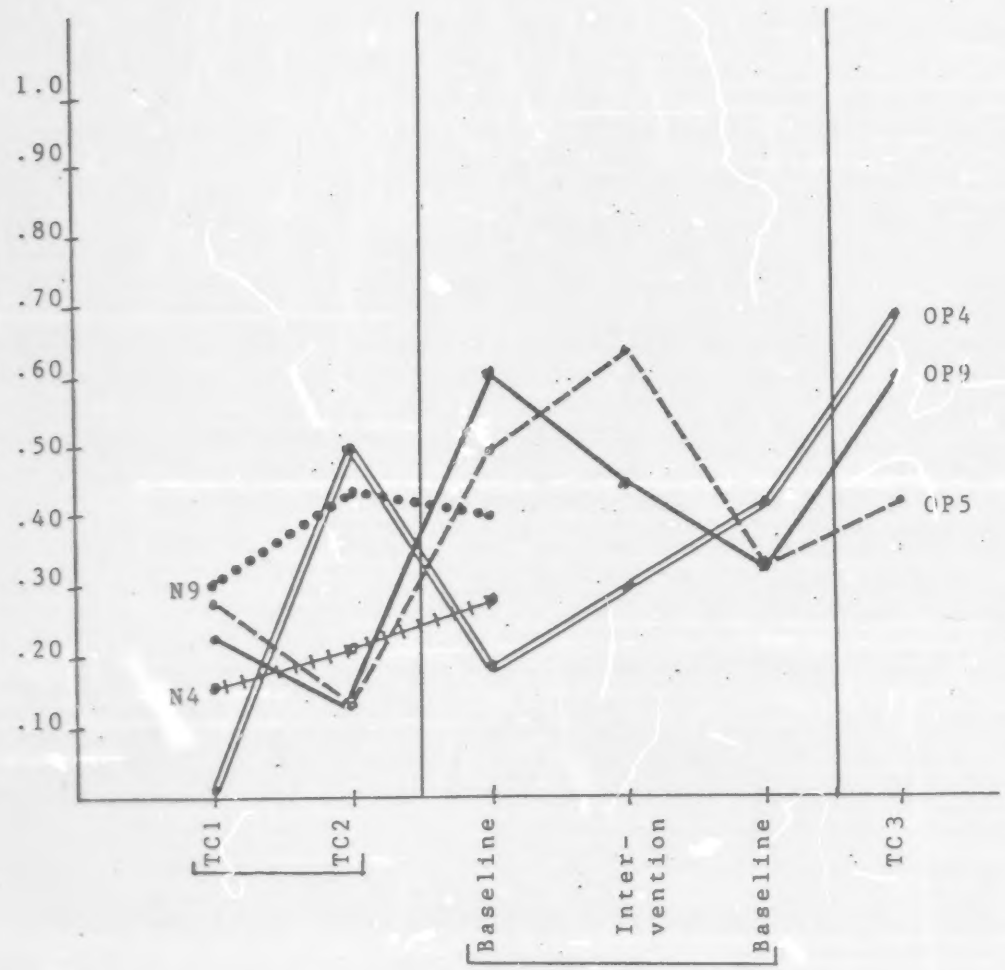


Figure 2

Proportion of Specific Conclusions

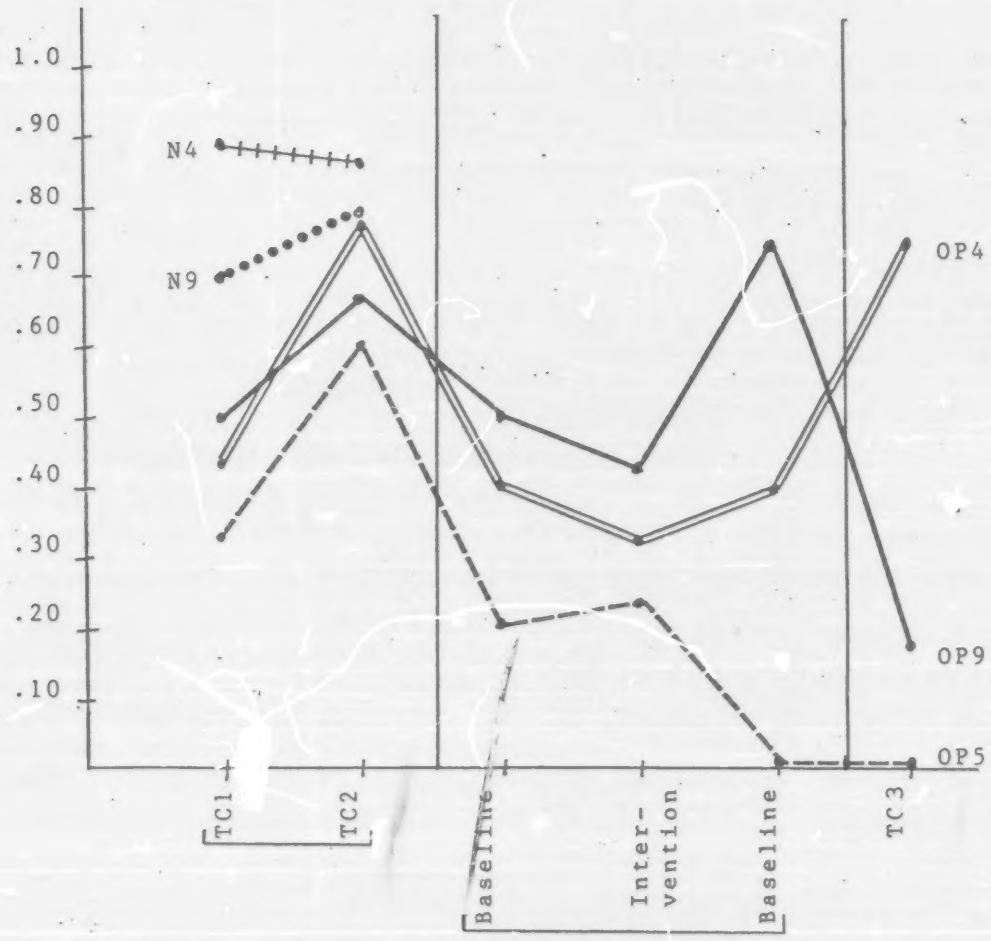


Figure 3

Proportion of Highly Articulated Conclusions

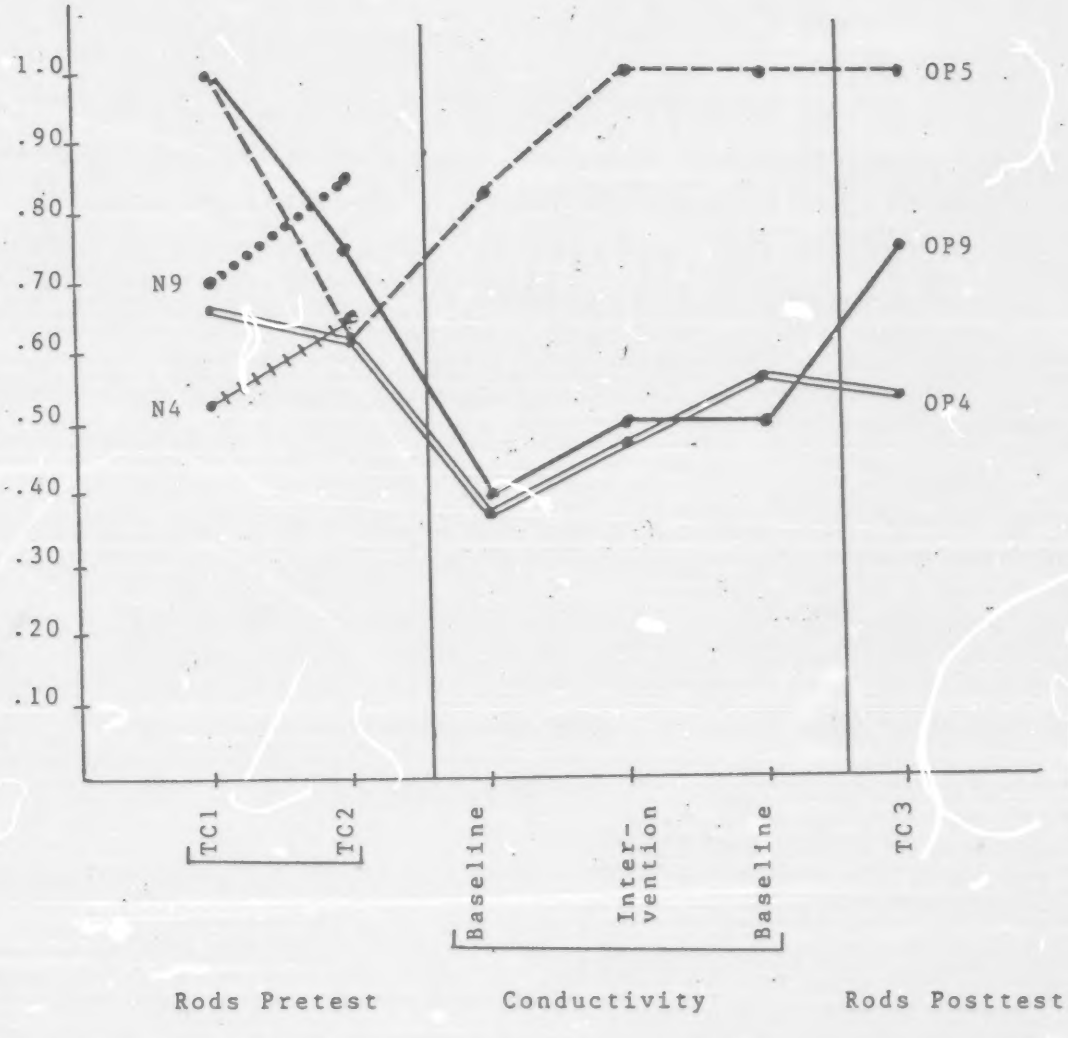


Figure 4

Proportion of Tests in which some kind of Explanatory Principle was used

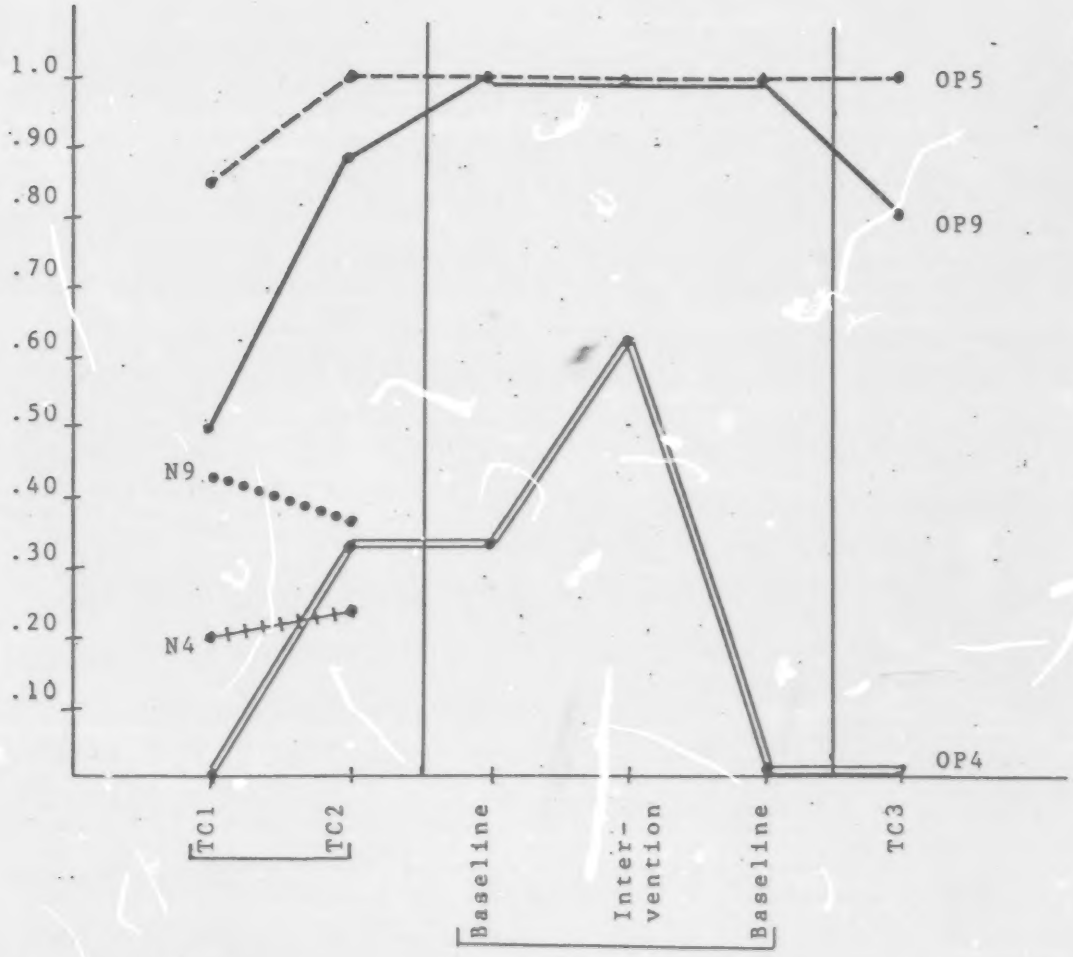


Figure 5

Proportion of Predictions with a
High Degree of Intentionality

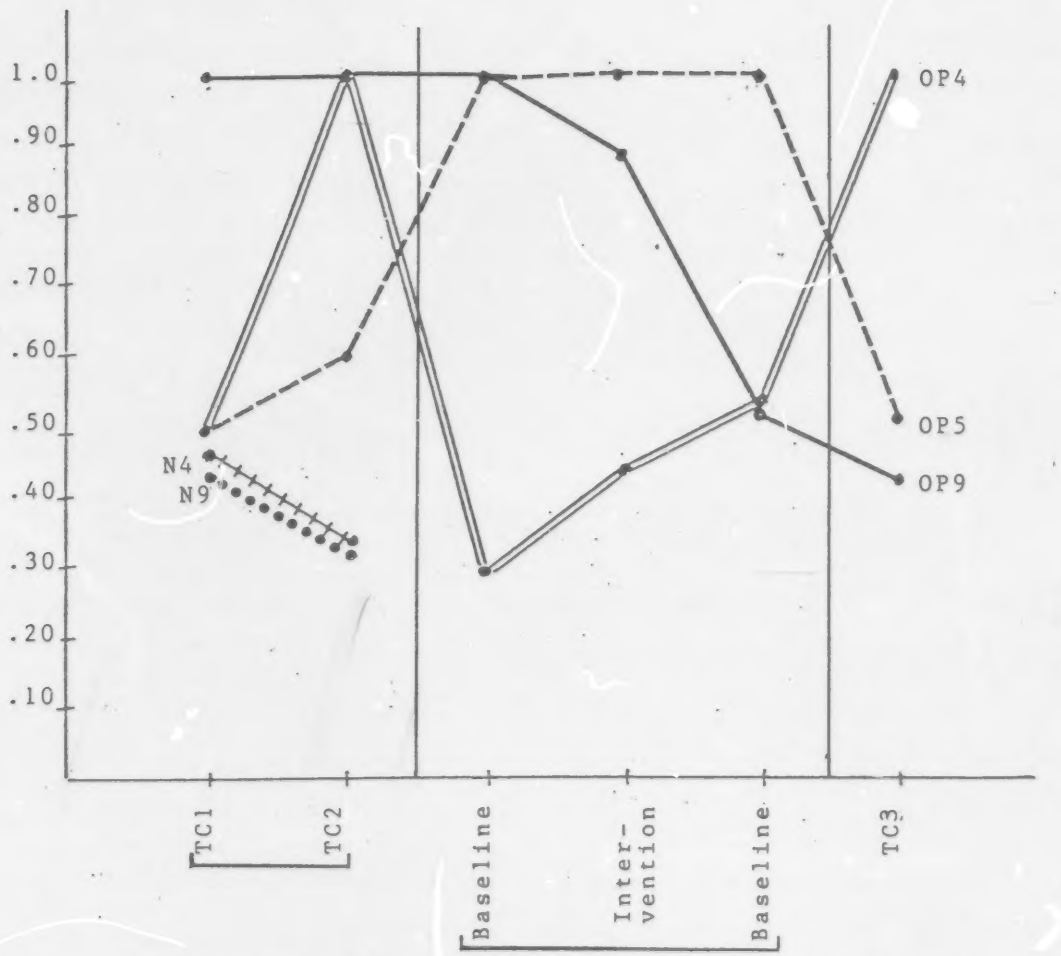


Figure 6

Proportion of Predictions that were Highly Explicit

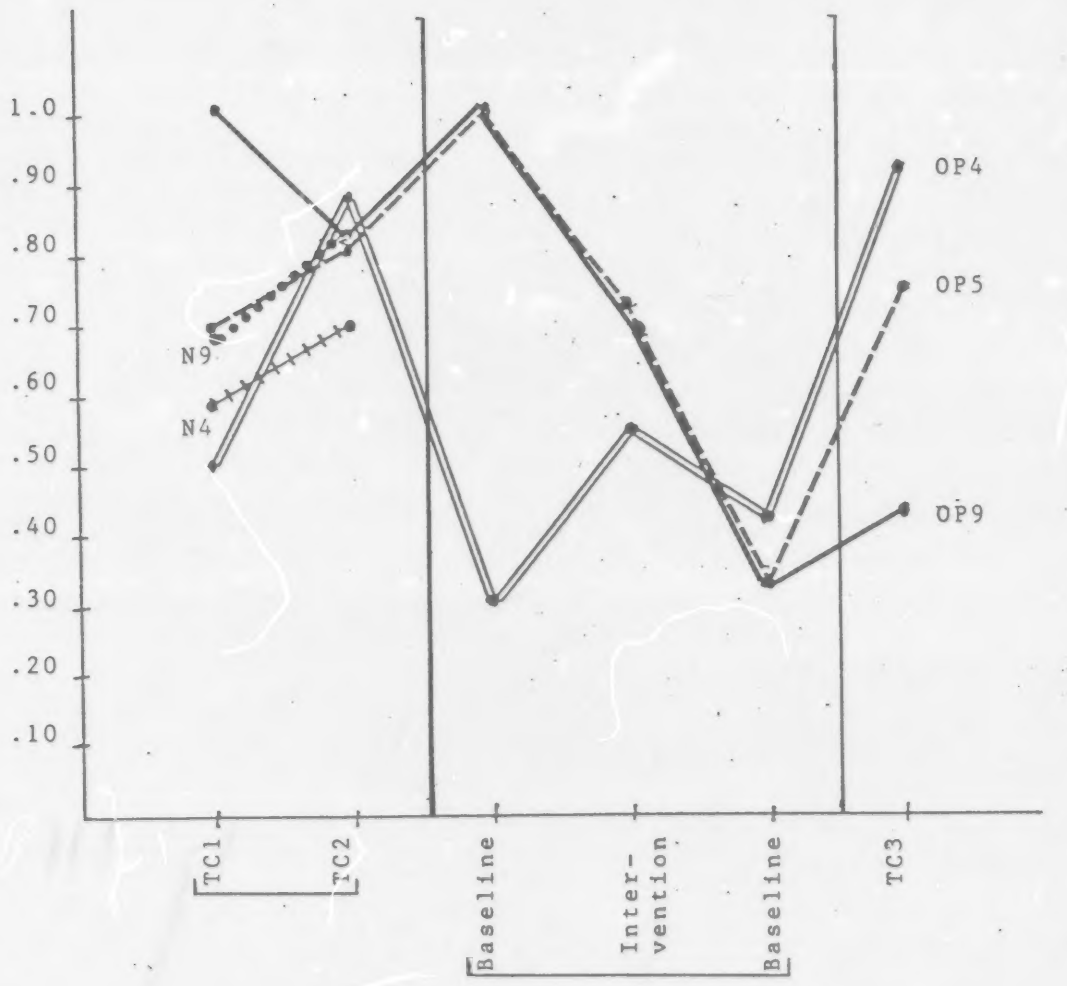
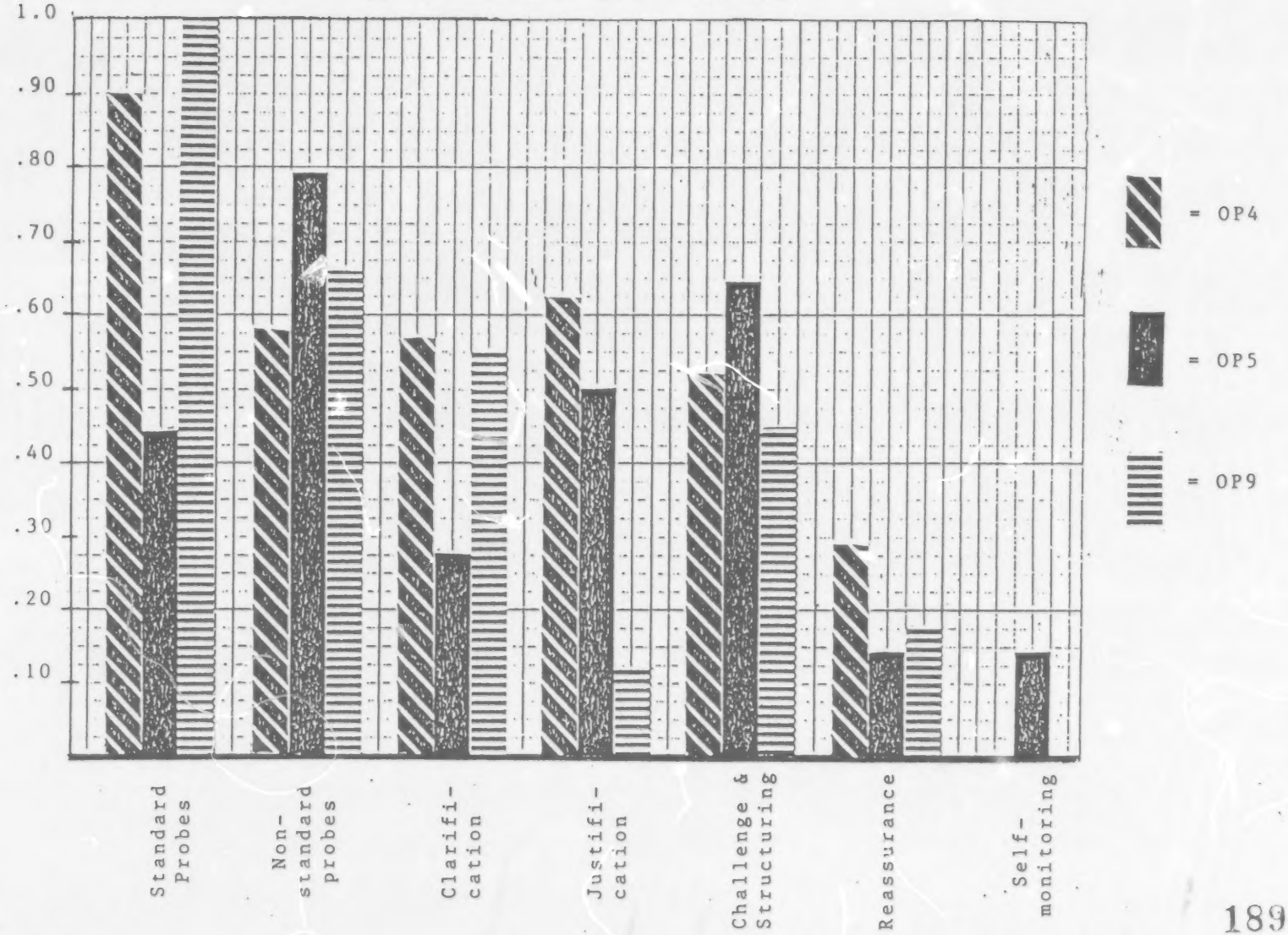


Figure 7

Proportion of tests in which each type of Examiner probe was used



Stone and Forman
7/84

APPENDIX A

Teacher Questionnaire

Dear LD Teacher:

_____ has agreed to participate in the Northwestern University Reasoning Study. It would help us if you could answer the following questions based on your knowledge of this student.

1. What do you consider to be his/her major difficulties (in order of severity? _____

_____

2. Does he/she have language problems? _____

If so, how would you describe his/her problems in this area? _____

_____

_____

Which of the following phrases could be used to describe his/her problems? (If inappropriate, leave blank.)

	Mild	Moderate	Severe	Unsure
Verbal memory problems	___	_____	_____	_____
Vocabulary comprehension problems	___	_____	_____	_____
Syntax comprehension problems	___	_____	_____	_____
Retrieval (word finding) problems	___	_____	_____	_____
Oral formulation problems	___	_____	_____	_____

3. Does he/she have reading problems? ___ _____

If so, how would you describe them? _____

_____

_____

Which of the following phrases could be used to describe his/her reading problems?

	Mild	Moderate	Severe	Unsure
Decoding problem	___	_____	_____	_____
Word or letter discrimination problem	___	_____	_____	_____
Sound-symbol correspondence problem	___	_____	_____	_____
Word comprehension problem	___	_____	_____	_____
Literal comprehension problem	___	_____	_____	_____
Main idea comprehension problem	___	_____	_____	_____

4. Does he/she have nonverbal problems? _____

If so, how would you describe them? _____

_____

Teacher Questionnaire Page Two

Which of the following phrases could be used to describe his/her nonverbal problems?

	Mild	Moderate	Severe	Unsure
Visual-motor integration	____	_____	_____	_____
Figure-ground problems	____	_____	_____	_____
Part-whole relations	____	_____	_____	_____
Spatial orientation/directionality	____	_____	_____	_____
Social perception problems	____	_____	_____	_____
Problems with size/distance estimation	____	_____	_____	_____

5. Does he/she have math problems? _____

If so, how would you describe them? _____

_____

_____

Which of the following phrases could be used to describe his/her math problems?

	Mild	Moderate	Severe	Unsure
Problems learning/remembering math facts	____	_____	_____	_____
Computation problems (whole numbers)	____	_____	_____	_____
Problems with measurement	____	_____	_____	_____
Problems with whole number concepts	____	_____	_____	_____
Problems with fraction concepts	____	_____	_____	_____
Difficulty with word problems	____	_____	_____	_____

6. Does he/she have problems with thinking skills? _____

If so, how would you describe them? _____

_____

_____

Which of the following phrases could be used to describe his/her problems?

	Mild	Moderate	Severe	Unsure
Problems with concept learning	____	_____	_____	_____
Difficulty with logical reasoning	____	_____	_____	_____
Difficulty judging relevant vs. irrelevant information	____	_____	_____	_____
Difficulty organizing a sequence or substeps	____	_____	_____	_____
Problems learning concepts in content areas	____	_____	_____	_____
Study skill problems	____	_____	_____	_____

APPENDIX B

SYNOPSIS OF ISOLATION OF VARIABLES TASK CODING SYSTEM

I. GOAL SETTING

A. NATURE OF OVER-ARCHING GOAL*

Purpose To characterize the subject's initial understanding of what he is to be done during the session on the basis of comments made during the first few tests.

Overview of Codes
1. Using rods
2. Trial and error/creating effects
3. Exploration/induction
4. Proof/deduction

B. INITIAL CONCEPTION OF PHENOMENON

1. Conception of outcome*

Purpose To characterize the subject's initial understanding of the dependent variable he is to investigate (e.g., differential bending or conductivity).

Overview of Codes
a. Binary
b. Continuous

2. Conception of factors involved*

Purpose To characterize the subject's initial assumptions concerning those aspects of the materials which are related to the outcome.

Overview of Codes
a. Number of factors
    1) Single
    2) Multiple-fused
    3) Multiple-separate
    4) Multiple-interacting
b. Nature of factors
    1) Discrete
    2) Continuous

---

* Codes marked with an asterisk have not yet been fully developed.

II. DATA GATHERING

A. ATTRIBUTE CODING

Purpose To assess the explicitness and appropriateness of the expressions used to refer to each standard variable.

Overview of Codes Each variable within a test will be coded as one of the following:
1. explicit, appropriate referring expression
2. explicit, inappropriate referring expression
3. presumed, appropriate referring expression
4. presumed, inappropriate referring expression
5. implied, appropriate referring expression
6. implied inappropriate referring expression
7. no referring expression, but variable is held constant
8. no referring expression, and variable is not held constant

B. UNIT OF INVESTIGATION*

Purpose To describe the unit of activity into which the subject has decomposed the task.

Overview of Codes
1. Isolated Rods
2. Pairs of rods
3. Sequences of pairs

C. ROD SCANNING AND SELECTING

Purpose To assess the degree to which the subject is reflective in his choice of rods and whether he is "idea" or "array" governed.

Overview of Codes Each test will be coded along each of the following dimensions:
- whether the rods selected were adjacent to one another
- how the rods were replaced in the array (e.g., replaced in exact position from which they were selected, separated from the array)
- number of seconds taken to select the first rod and the second rod
- the number of times the subject picks a rod to be used in the test but then puts it back into the array because he wishes to use a different rod

We will also note across the entire session whether the subject proceeds in a consistent left-to-right or right-to-left manner.

D. PREFERENCE FOR UNCONFOUNDED TESTS

1. Verbal measure of number of unconfounded tests
Purpose To assess explicit use of a control-of-variables strategy (based on rod pairs chosen and the subject's utterances).

---

* Codes marked with an asterisk have not yet been fully developed.

2. Nonverbal measures

   a. Number of uncredited unconfounded rod pairs
      Purpose To credit the subject for selection of rod pairs varying on only one dimension in the absence of any explicit verbalizations concerning the conclusions drawn.
   b. Number of valid confounded tests
      Purpose To credit the subject for selecting rod pairs in which the confounding variable(s) works against the conclusion drawn.

E. SEQUENCING OF ACTIONS WITHIN TESTS*

Purpose To note extreme variations across subjects (and across tests for a given subject) in the order in which component activities (e.g. attaching weights, drawing conclusions, etc.) are carried out.

F. INCORPORATION OF NEW DATA*

Purpose To characterize the manner in which the subject integrates new information into his ongoing store of observations.

Overview of Codes
1. Ignored
2. Distorted
3. Integrated

G. VERIDICALITY OF OBSERVATIONS*

Purpose To rate the degree to which the subject observes materials and events accurately.

H. META-STRATEGIC COMMENTS*

Purpose To capture the frequency and nature of the subject's comments about his own activity.

Overview of Codes
1. On-line narration
   a. Frequency
   b. Relation to action
      1) Temporal
      2) Referential
2. Explicit reference to strategy

---

* Codes marked with an asterisk have not yet been fully developed.

III. REASONING PROCESS

A. REASON FOR SUCCESSIVE TESTS MADE*

Purpose To characterize the subject's reasons for the tests he conducts and the sequence in which he conducts them, as inferred from verbalizations.

Overview of Codes
1. Array-driven
2. Attribute or effect driven
3. Theory-driven
4. Insufficient information

B. NATURE OF TESTS OF EACH VARIABLE

Purpose To assess how each target variable is tested during the task as a whole.

Overview of Codes Each variable will be coded as one of the following:
1. No spontaneous or prompted tests
2. Prompted tests only
3. A single test
4. Multiple tests

C. USE OF PRIOR INFORMATION*

Purpose To rate the degree to which the subject appears to make use of information from earlier tests in drawing conclusions from later tests.

Overview of Codes
1. None
2. Use of knowledge
3. Relevant information from past tests/outcome disregarded
4. Explicit reference to prior tests/outcomes made
5. Implicit reference to prior tests/outcomes made

D. PREDICTIONS

Purpose To assess intention prior to actual testing and the explicitness of the intention.

Overview of Codes Each variable that has a test notation next to it will be coded along the following two scales:

Intention	Explicitness
0. Conclusion only	0. Conclusion only
1. No prediction	1. No variable specified
2. Rod description only	2. Variable partially specified
3. Sense of intention only	3. Maximally explicit
4. Empirically based predictions	
5. Hypothetical prediction	

---

* Codes marked with an asterisk have not yet been fully developed.

E. NATURE OF CONCLUSIONS DRAWN

1. Nature of Comparisons Made*

Purpose To note the data base on which each conclusion is made.

Overview of Codes
a. No comparison
b. Comparisons made between 2 rods in stand
c. Comparisons made betwen this test and previous tests
d. Comparisons made between one rod in stand and previous rods or tests
e. Unclear comparisons
f. Comparisons between rods in stand and hypothetical test
g. Are the above comparisons parallel or not?

2. Degree of Articulation of Relationship Between Attributes and Outcomes

Purpose To characterize the conceptualization of the relationship between causal variables and outcome.

Overview of Codes
0. Not applicable
1. No relation noted
2. Partial correspondence
3. Complete correspondence
4. Continuous functional relationship

3. Nature and Adequacy of Explanatory Principle

Purpose To assess the conceptual sophistication of a subject's explanations for the target phenomenon.

Overview of Codes
0. Not applicable
1. No explanatory principle offered
2. Attributes identified as causal agents
3. Tautological explanation
4. Inferred construct

4. Generality of Conclusions

Purpose To assess the degree to which a subject's conclusions extend beyond the test at hand.

Overview of Codes
0. Insufficient information
1. Conclusions directed to specific rods
2. Conclusions directed to a subset of the array
3. Conclusion general

---

* Codes with an asterisk have not yet been fully implemented.

IV. EXAMINER GUIDANCE

A. REQUESTS FOR GUIDANCE*

Purpose To note number and nature of a subject's requests for guidance.

B. NECESSITY FOR PROBES*

Purpose To note the extent to which the examiner was forced to use probes to elicit tests and/or conclusions from a subject.

C. COMPLIANCE WITH PROBES

Purpose To determine the percentage of times the subject complies with the examiner's request to test a particular variable.

Overview of Codes The total number of compliances and noncompliances will be summed across the variables for each segment of the sessions.

---

* Codes marked with an asterisk have not yet been fully developed.

APPENDIX C

Teacher Rating Scale for the Bending Rods Task

Purpose

The purpose of the Teacher Rating Scale is to give LD practitioners an objective tool for refining and standardizing their clinical judgments in the diagnosis and remediation of specific reasoning and problem solving deficits. You will be asked to observe adolescents engaged in a reasoning task, the bending rods, and to rate their performance according to coding procedures outlined in this manual.

Description of the Bending-Rods Task

The bending-rods task focuses on a single reasoning strategy, the control of variables, which is not only applicable to everyday life but is also a key ingredient in the assessment of the transition to the stage of formal operations, as described in Piagetian theory. Formal operations, which develop during early adolescence, are said to provide a foundation for the mastery of the abstract concepts and critical thinking skills necessary at higher educational levels and for making important real-world decisions.

The subject's task is to determine how certain variables affect the bending of rods which the subject inserts into a stationary stand and causes to bend by attaching weights to the free end of each rod. Some rods bend more than others, and the subject is asked to figure out "all the things that matter for bending" by comparing the relative bending of pairs of rods. The materials for the task consist of an array of twelve different rods (plus two demonstration rods), each permanently mounted in a short, cylindrical base; a stand into which pairs of rods can be inserted for testing; and weights which can be attached to the free ends of the rods to make them bend. Five variables (of which only four "matter for bending") are apparent in the materials. The array of twelve rods includes various combinations of the following three variables:

length:	20 cm	40 cm	60 cm
diameter:	.32 cm	.48 cm	.64 cm
material:	steel	wood	plastic

The fourth variable is the two pairs of weights, one set obviously heavier (and larger) than the other; the subject may choose to attach either similar weights (both light or both heavy) or dissimilar weights (one of each) for testing any pair of rods. The fifth variable, which is irrelevant for bending, is the material from which the base is made; some are plastic and some are chrome.

Depending on which rods and weights the subject selects for comparison, any given test may be said to be confounded (not controlled) or unconfounded (controlled) with respect to a particular variable. For example, in an unconfounded test of the effect of diameter on bending, the rod pair would be similar in all respects except diameter (same length, same material, same base, same weights attached). A subject who consistently avoids confounded

tests in the bending-rods task is presumed to have mastered and internalized the control-of-variables strategy.

Experimental Procedure

Each experimental session is composed of several parts, of which you will be coding only the first. In this part of the session, the experimenter first uses a pair of rods (not part of the array of twelve) to demonstrate how the rods are inserted into the stand, how the weights are attached for a test of bending, and how some rods bend more than others. [The demonstration constitutes an unconfounded test for length -- that is, all other variables (diameter, material, weights attached, and base) are "controlled" in that they are the same for both rods, but this is not pointed out to the subject]. The experimenter then asked the subject to figure out all the things that make some rods bend more than others by choosing pairs of rods from the array of twelve and testing them.

For each test, the examiner first asks, in effect, "Why did you pick those two rods?" This is called the "Why pick" question, and the subject's response is called a "reason." The examiner subsequently asks "What did you learn about bending from trying those two rods?" This is called the "What learn" question, and the subject's response is called a "conclusion." Part of the coding attempts to analyze the subject's verbal response to these two questions.

After these "spontaneous" tests, the experimenter asks the subject to summarize what he or she has learned "so far" about bending. The experimenter then elicits several "prompted" tests by asking questions about specific variables.

Coding Procedure

You will use a worksheet (to which you should refer as you read the following procedures) to code various aspects of the subject's behavior for each test he or she conducts during the "spontaneous" portion of the experimental session. The various behaviors to be coded are arranged, left to right, in roughly the same sequence in which the behavior or response occurs during each test. The aspects of behavior to be coded include

1. the manner of rod selection;
2. whether the test was confounded or unconfounded for a particular variable;
3. the nature of the reasons given for the selection of a particular pair of rods;
4. the nature of the conclusions drawn after each test;
5. a characterization of any explanations offered by the subject for why some rods bend more than others; and
6. the variables mentioned in the subject's verbal response.

Specific guidelines for coding each aspect are given below.

Rod Selection:

The manner in which the subject (S) selects the rods for each test should be rated as either "random" or "thoughtful" by putting a check in the appropriate column.

Random: S selects rods immediately, without appearing to contemplate choice. This category should also include selections that are determined by the array rather than by the characteristics of the various rods, as when, for example, the subject chooses the rods in the same order in which they are arrayed before him, starting at one end and proceeding down the line.

Thoughtful: S exhibits indications of reflectiveness, such as a 2- or 3-second pause for thinking, one or more false starts (picks up a rod, then replaces it and selects another), or an examination of the characteristics of the rods.

Decisions regarding thoughtfulness of rod selection should be made after observing both rod choices for a given test. If one rod is selected quickly and the other thoughtfully, rate the selection as "thoughtful."

Confounded/Unconfounded:

The attributes of the rods and weights selected for each test should be recorded on the worksheet. The following conventions will be used to refer to the variables represented in any given test conducted by the subject:

length: 20, 40, or 60 (short, medium, or long)

diameter: 2, 3, or 4 (thin, medium, or thick; that is, 2/16, 3/16 or 4/16 of an inch)

material: S, W, or P (steel, wood, or plastic)

weight: light -- a circle around the first variable (length)
heavy -- a circle around the second variable (diameter)

base: chrome -- no mark
plastic -- a line under the third variable (material)

Example: 20(2)S : 40(2)S

The rod attributes are always given in the same order: length, diameter, and material. The example above refers to a test in which the bending of a short, thin, steel rod with a chrome base is compared to that of a medium, thin, steel rod with a plastic base by attaching the heavier weights to both rods.

A determination is to be made as to whether each test is confounded or unconfounded with respect to a particular variable, based not only on the actual rods chosen but also on what S says about the test. For example,

The type of base is irrelevant for bending. However, many subjects hypothesize that one type of base allows the rods to bend more than the other, and therefore the base must be taken into account in deciding whether a test is confounded or not.

Two-step Coding Procedure:

First step: Inspect the rods selected by the subject and make one of the following entries in the "confounded or unconfounded" column:

C	The test is confounded by two or more variables other than the base.
U-variable	"U-length," for example, indicates that the test is an unconfounded test of the effect of length.
C-variable	Draw a line under the "C" if the test is confounded by base alone. For example, 20 4 S : 20 3 S would be coded "C-diameter."

Second step: Two aspects of the coding must then be clarified:

a. For unconfounded tests of a particular variable, you must verify that the subject actually intended to test that variable. If S's response to the "Why pick" question includes mention of the relevant variable, let the first-step coding stand. If no reference is made to the variable, write "[NR]" after the original coding.

   Note: S's verbal response regarding an unconfounded test may include other variables as part of a referring expression. For example, "I was testing if the short metal one was stiffer" would be an acceptable "verification" of an unconfounded test of material.

b. You must decide whether the tests confounded by base alone (C) are actually confounded or unconfounded from the subject's viewpoint. If the subject has established (or assumed) that the base does not matter for bending, write "[U]" after the original notation. If the subject has established (or assumed) that the base does matter, write "[C]" after the original notation.

   Notes:

   Some subjects may change their minds about the role of the base several times in the course of the testing; each test should be coded according to the subject's current opinion.

   If it is not possible to tell what the subject's opinion is, assume that the base is a relevant, confounding variable and write "[C]" after the original notation.

Reason:

The purpose of this code is to categorize the subject's reasons for rod selection in terms of his or her apparent intent to test a hypothesis. The subject's response to the "Why pick" question should be classified in one of the four categories described below by putting a check in the appropriate column on the worksheet.

1. None offered.

   Examples: I don't know.
   They're next in line.
   Same reason as the rest.

2. Description of variables only.

   Examples: Different lengths.

   It, ah, they're the same size, but one's heavier, the weights are heavier.

   This one's thicker.

   Thought they were the same. (Presumably refers to rod characteristics.)

3. Intention to compare bending only (no mention of variables).

   Examples: To see which one would go down more.

   Wanted to see if that one bends.

   Figured they'd be equal. (Presumably refers to bending.)

4. Intention to compare effect of variables on bending.

   Statements in this category refer both to rod (and weight) characteristics and to the intention to note relative bending. Typically, a strong sense of intent to compare bending is conveyed through phrases such as "I wanted to see . . .," and "I was comparing . . . ."

   Examples: I was seeing if this one was the metal one and this a wood one. See which one pulled out longer.

   Because they're about the same length. And this one's made out of a thicker piece of fiberglass than the metal rod. And, ah, I wanted to see how much the thickness made a difference.

Variables mentioned in reason:

The specific variables actually mentioned in response to the "Why pick?" question should be indicated by circling the appropriate letters on the worksheet.

L = length
D = diameter ("thickness")
M = material
B = base
W = weights

Note: It is not necessary that the subject use standard terms in referring to the variables. E.g., "skinny" may refer to diameter, "bottom" to the base, "sinkers" to the weights. "Bigger" (or similar expressions, including "size") will usually refer to length, but occasionally it will be clear from the context that such an expression refers to diameter. Nonverbal indications (such as pointing gestures) and responses made in other tests may be relied upon to clarify which variable is meant. If it is not possible to determine which variable an expression refers to, put a question mark in the "variables mentioned" column and note the ambiguous expression in the "comments" column.

Conclusions:

Conclusions are those statements made in response to the "What learn" question. Each conclusion will be coded for generality according to whether the conclusion (the subject's description of the outcome of the test) seems to pertain only to the pair of rods tested or whether it seems to refer to the relative bending of all possible rod pairs.

1. Single-test conclusions.

   Examples: This one bent more.
   The long one bent more than the short one.

2. Generalized conclusions.

   Statements in this category may include superordinate terms such as length or thickness, and they seem to focus on whether a particular attribute makes a difference for bending in general, rather than solely for the particular pair of rods in the stand.

   Examples: Long ones bend more than short ones.
   The thinner the rods are, the more they bend.

Note: Sometimes the "What learn" question is not asked, either because the subject drew a conclusion about bending in his response to the "Why pick" question or because of experimenter oversight.

Example: I wanted to see if the long rod would bend more than the short rod, and it did.

This statement should be coded as both a Category 4 "reason" and a Category 1 "conclusion."

When one or the other of the probe questions is not asked, you should nevertheless complete the coding as if both questions had been asked, insofar as it is possible to do so. When there is no codable response, draw a horizontal line through the appropriate boxes on the worksheet to indicate that the lack of categorization is not a coder oversight.

Explanations:

Although no probe question is asked to elicit the subject's ideas about why some rods bend more than others, such "explanations" sometimes arise in response to the "What learn" question. The coding distinguishes between explanation based on the rod and weight characteristics and explanation which evokes some other scientific construct (usually "gravity").

1. None offered.

   The conclusion describes the outcome of the test ("Long ones bend more"), but no cause is stated.

   Statements such as "The thicker it is, the less it goes down" are considered to be non-causal observations and are therefore included in this category.

2. Rod and weight attributes identified as causal agents.

   a. Explicit markers:

      This category includes all statements in which linguistic markers (such as because, so, therefore, and makes it) point to a causal relationship between the rod and weight characteristics and the fact that one rod bends more than another.

      Examples: Length makes it heavy.
      They bend more because they're skinny.
      It's more sturdy so it goes down more.

      Note: The first two statements would also be coded as Category 2 conclusions (generalized); the last statement would also be coded as a Category 1 conclusion (single-test).

   b. Implied causation:

      This category includes statements in which the description of the outcome implies a causal relationship between rod attributes and bending.

      Examples: This steel one's sturdier material.
      Short ones are less flexibler than long ones.

      A test for such an implication is whether it seems reasonable to add "and so it bends more (or less)" to the subject's actual response. If S had merely said, "This steel one didn't bend," or "Short ones bend less than long ones," you would code his response as a Category 1 explanation (none offered).

Note: Terms such as the following are considered to be variations on the rod and weight attributes and are therefore included in this category:

weak	weakness
strong	strength
flexible	"flex," flexibility
stiff	stiffness
sturdy	sturdiness

3. Other scientific constructs.

   Include in this category only explanations which refer to constructs which are not directly related to the rod and weight attributes, such as "gravity" or any reference to "molecules" or "density."

Variables mentioned in conclusion and explanation:

Any variables not already coded that are mentioned in response to the "What learn" question should be circled in the "variables mentioned" column.

Comments:

Use this part of the worksheet to indicate any special circumstances which seem to make the coding difficult or arbitrary. Ambiguous and unusual referring expressions should be noted here, as well as any misperceptions on the part of the subject (for example, S refers to one of two identical rods as being "longer"). When possible, write S's responses verbatim.

Summary

As you observe the session, you must gather the information listed below.

1. During rod selection, observe whether "random" or "thoughtful."
2. If test is unconfounded, look for evidence of intention to test the controlled variable.
3. If test is confounded by base alone, decide whether the test is actually confounded or unconfounded from the subject's viewpoint.
4. Analyze the response to the probe questions, deciding:
   a. what type of reason was supplied.
   b. whether the conclusion was specific or generalized.
   c. whether an explanation was offered and, if so, what type.
   d. what variables have been named.

test no.	rods selected: (L) light, (D) heavy, M plastic	rod selection: random	rod selection: thoughtful	confounded or unconfounded	reasons "Why pick?": none 1	reasons: var. only 2	reasons: bend only 3	reasons: effect of var. 4	variables mentioned	conclus. "what learn": single test 1	conclus.: generalized 2	explanations: none 1	explanations: rod attributes: exp. 2a	explanations: rod attributes: imp. 2b	explanations: other 3	comments
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
									L D M W B							
					/				L D M W B							

TRS Coding Form

Stone and Forman
7/84

BEST COPY AVAILABLE

APPENDIX D

Dissemination

The findings of the project are being disseminated through several means. The prinicpal investigators gave presentations based on the findings at two national professional conventions in 1983, the Council for Learning Disabilities and the Society for Research in Child Development.. Proposals for presentations of additional aspects of the data have been submitted to the American Educational Research Association and the Society for Research in Child Development.

The findings have been shared with LD practicioners via presentations to the Special Education staffs of several local schools and at a workshop as part of a special symposium organized by the Orton Dyslexia Society. Also, all schools which participated in the project were sent a twenty page summary of the findings.

Currently, the principal investigators are working with former research assistants to write summaries of the findings suitable for publication in appropriate professional journals. The findings are alos discussed in a review chapter on the problem solving skills of LD children written by Stone and Michals which will appear in a forthcoming handbook on learning disabilities.